US009349478B2

(12) **United States Patent**
Yuan et al.

(10) **Patent No.: US 9,349,478 B2**
(45) **Date of Patent: May 24, 2016**

(54) **READ WITH LOOK-BACK COMBINED WITH PROGRAMMING WITH ASYMMETRIC BOOSTING IN MEMORY**

(71) Applicant: **SanDisk Technologies Inc.**, Plano, TX (US)

(72) Inventors: **Jiahui Yuan**, Fremont, CA (US); **Yingda Dong**, San Jose, CA (US); **Charles Kwong**, Redwood City, CA (US); **Hong-Yan Chen**, Sunnyvale, CA (US); **Liang Pang**, Fremont, CA (US)

(73) Assignee: **SanDisk Technologies Inc.**, Plano, TX (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 120 days.

(21) Appl. No.: **14/500,660**

(22) Filed: **Sep. 29, 2014**

(65) **Prior Publication Data**

US 2016/0093390 A1 Mar. 31, 2016

(51) **Int. Cl.**

G11C 16/34	(2006.01)
G11C 16/26	(2006.01)
G11C 16/04	(2006.01)
G11C 16/10	(2006.01)
G06F 11/10	(2006.01)
G11C 29/52	(2006.01)

(52) **U.S. Cl.**

CPC ***G11C 16/3427*** (2013.01); ***G06F 11/1068*** (2013.01); ***G11C 16/0466*** (2013.01); ***G11C 16/0483*** (2013.01); ***G11C 16/10*** (2013.01); ***G11C 16/26*** (2013.01); ***G11C 16/3459*** (2013.01); ***G11C 29/52*** (2013.01)

(58) **Field of Classification Search**

CPC G11C 16/10; G11C 16/16; G11C 16/26; G11C 16/3418; G11C 16/3427; G11C 29/026; G06F 11/1072

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,292,476 B2 * 11/2007 Goda G11C 16/0483 365/185.17
7,440,318 B2 * 10/2008 Fong G11C 16/3418 365/185.02
7,864,582 B2 * 1/2011 Lee G11C 16/16 365/185.17

(Continued)

OTHER PUBLICATIONS

International Search Report & the Written Opinion of the International Searching Authority sated Nov. 17, 2015, International Application No. PCT/US2015/048001.

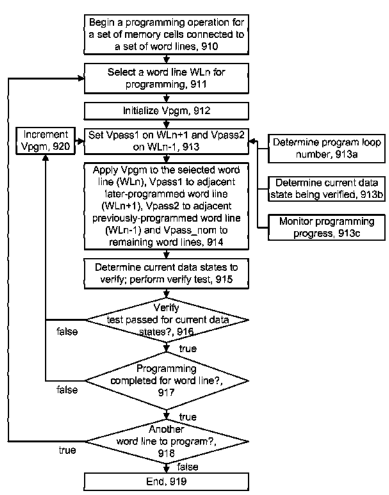
(Continued)

Primary Examiner — Shelly A Chase
(74) *Attorney, Agent, or Firm* — Vierra Magen Marcus LLP

(57) **ABSTRACT**

A read operation compensates for program disturb when distinguishing between an erased-state and a lowest programmed data state, where the program disturb is a function of the data state of an adjacent, previously-programmed memory cell on a common charge-trapping layer. A programming operation avoids program disturb of the programmed data states by using asymmetric pass voltages. Before reading the memory cells on a selected word line (WLn), the memory cells on the adjacent, previously-programmed word line (WLn−1) are read. The read operation for WLn uses multiple read voltages—one for each data state on WLn−1, and one of the read results is selected based on the data state of the adjacent memory cell. Other read operations distinguish between each pair of adjacent programmed data states using a read voltage which is independent of the data state of the adjacent memory cell.

21 Claims, 18 Drawing Sheets



(56) **References Cited**

U.S. PATENT DOCUMENTS

7,898,864	B2 *	3/2011	Dong	G11C 11/5628 365/185.03
8,051,240	B2	11/2011	Dutta et al.	
8,456,918	B2 *	6/2013	Oh	G11C 16/10 365/185.02
2009/0086542	A1	4/2009	Lee et al.	
2011/0305088	A1	12/2011	Huang et al.	
2012/0113723	A1	5/2012	Sarin et al.	
2012/0240011	A1	9/2012	Helm et al.	
2014/0160866	A1	6/2014	Shur et al.	

OTHER PUBLICATIONS

U.S. Appl. No. 14/322,055, filed Jul. 2, 2014.
U.S. Appl. No. 14/481,304, filed Sep. 9, 2014.

* cited by examiner

Fig. 1A

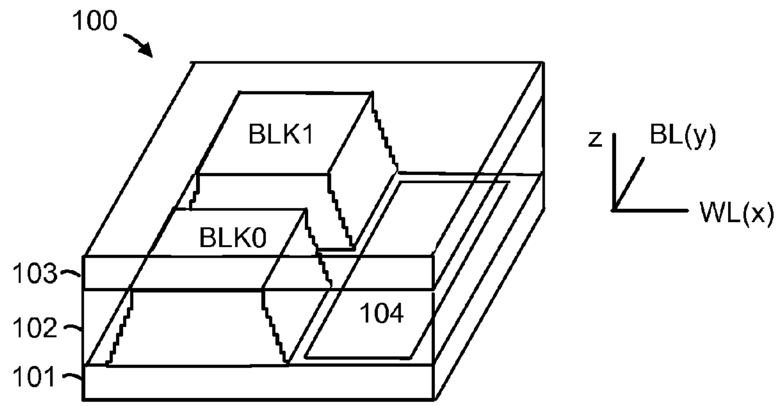
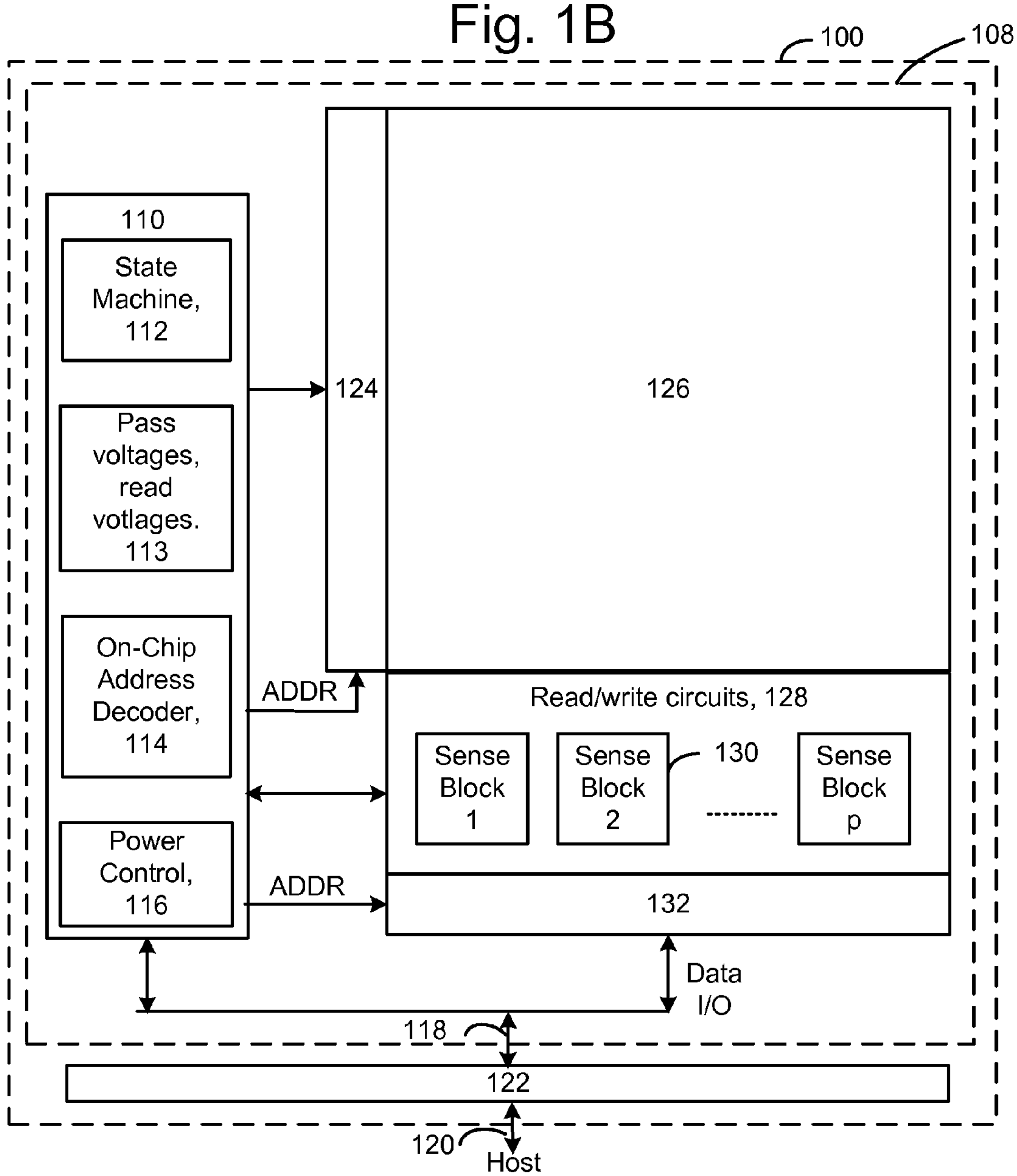

100
BLK1
BLK0
103
102
101
104
z
BL(y)
WL(x)

Fig. 1B


100
108
110
State Machine, 112
Pass voltages, read votlages. 113
On-Chip Address Decoder, 114
Power Control, 116
124
126
ADDR
Read/write circuits, 128
Sense Block 1
Sense Block 2
130
Sense Block p
ADDR
132
Data I/O
118
122
120
Host

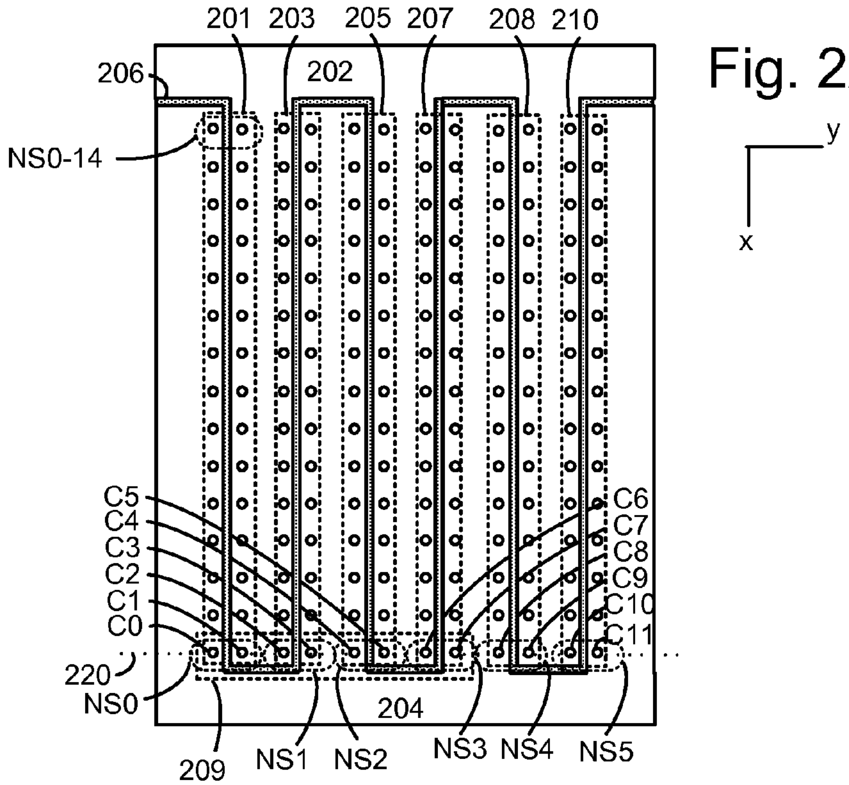

201
203
205
207
208
210
206
202
NS0-14
y
x
C5
C4
C3
C2
C1
C0
C6
C7
C8
C9
C10
C11
220
NS0
204
209
NS1
NS2
NS3
NS4
NS5

Fig. 2A

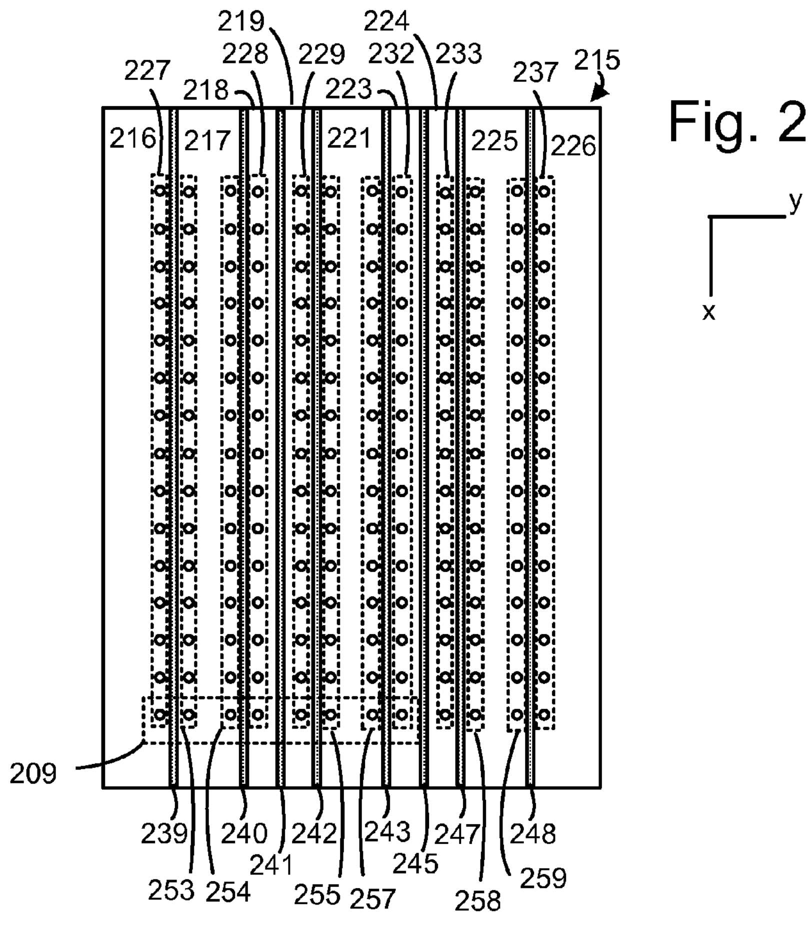
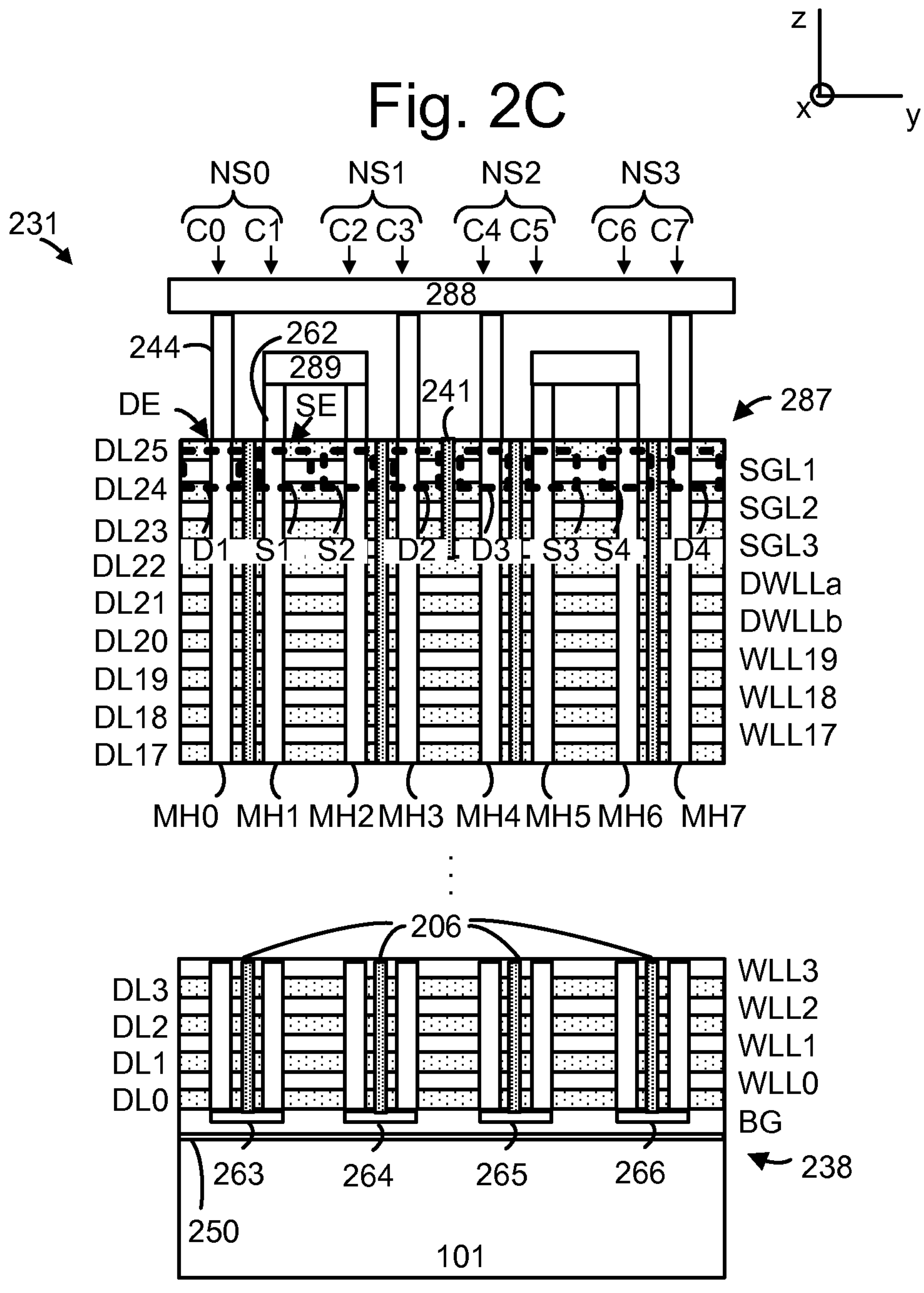

219
224
228
229
232
233
215
227
218
223
237
216
217
221
225
226
y
x
209
239
240
242
243
247
248
241
245
253
254
255
257
258
259

Fig. 2B

Fig. 2C



z
x
y
NS0
NS1
NS2
NS3
C0 C1
C2 C3
C4 C5
C6 C7
231
288
244
262
289
241
287
DE
SE
DL25
DL24
DL23
DL22
DL21
DL20
DL19
DL18
DL17
SGL1
SGL2
SGL3
DWLLa
DWLLb
WLL19
WLL18
WLL17
D1 S1 S2 D2 D3 S3 S4 D4
MH0 MH1 MH2 MH3 MH4 MH5 MH6 MH7
206
DL3
DL2
DL1
DL0
WLL3
WLL2
WLL1
WLL0
BG
263
264
265
266
238
250
101

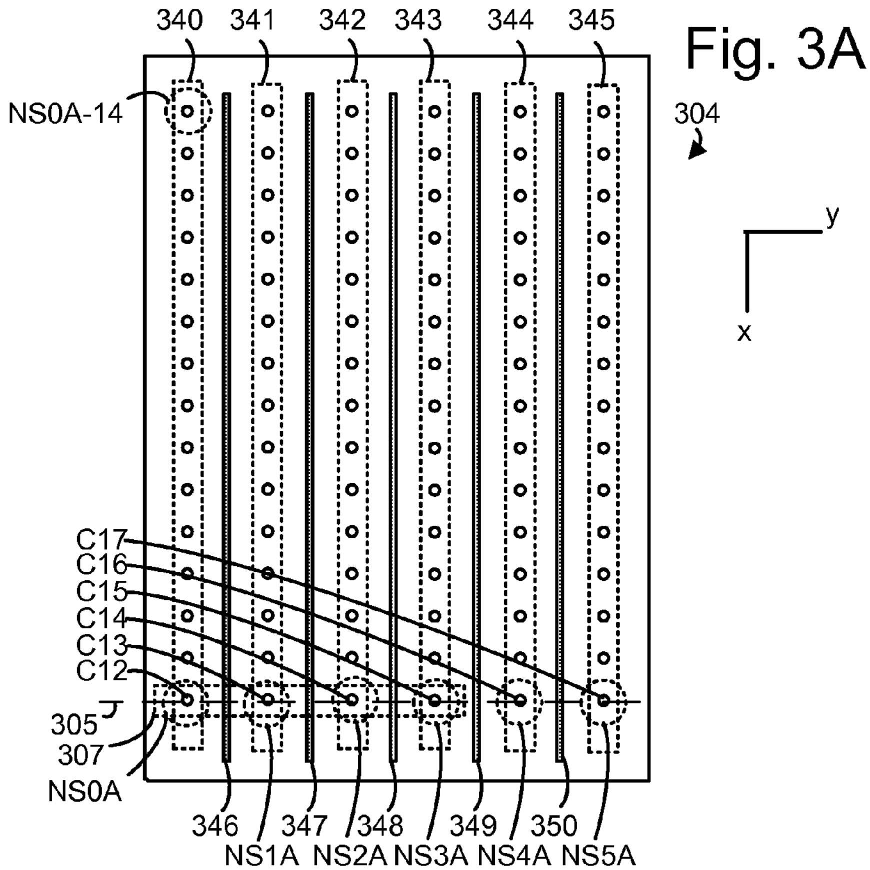

340
341
342
343
344
345
304
y
x
NS0A-14
C17
C16
C15
C14
C13
C12
305
307
NS0A
346
347
348
349
350
NS1A
NS2A
NS3A
NS4A
NS5A

Fig. 3A

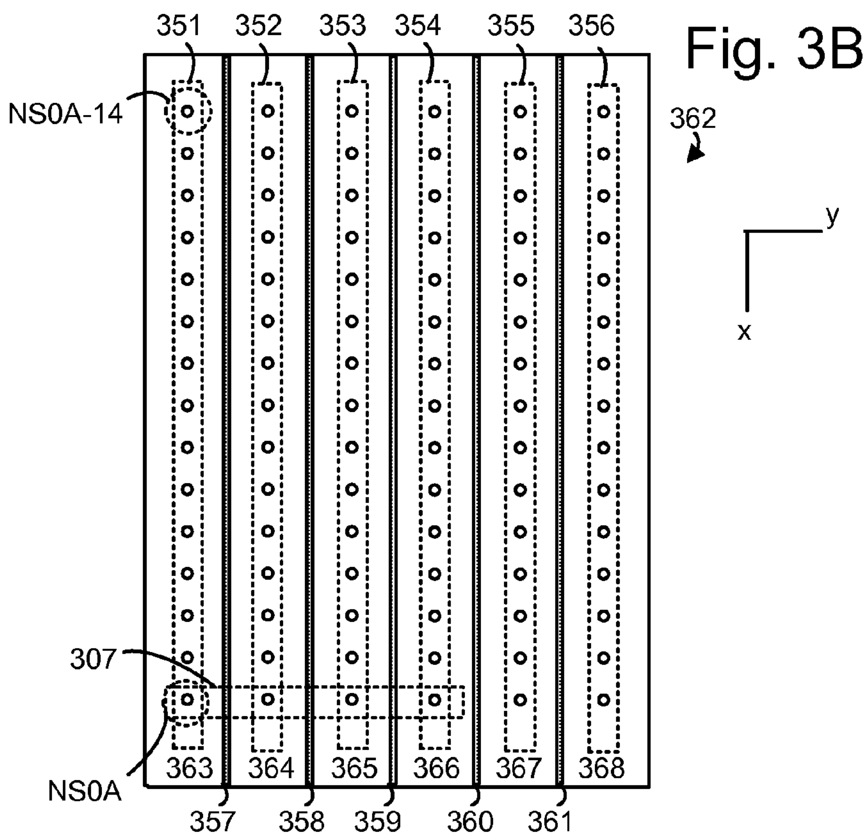

351
352
353
354
355
356
362
y
x
NS0A-14
307
NS0A
363
364
365
366
367
368
357
358
359
360
361

Fig. 3B

Fig. 3C

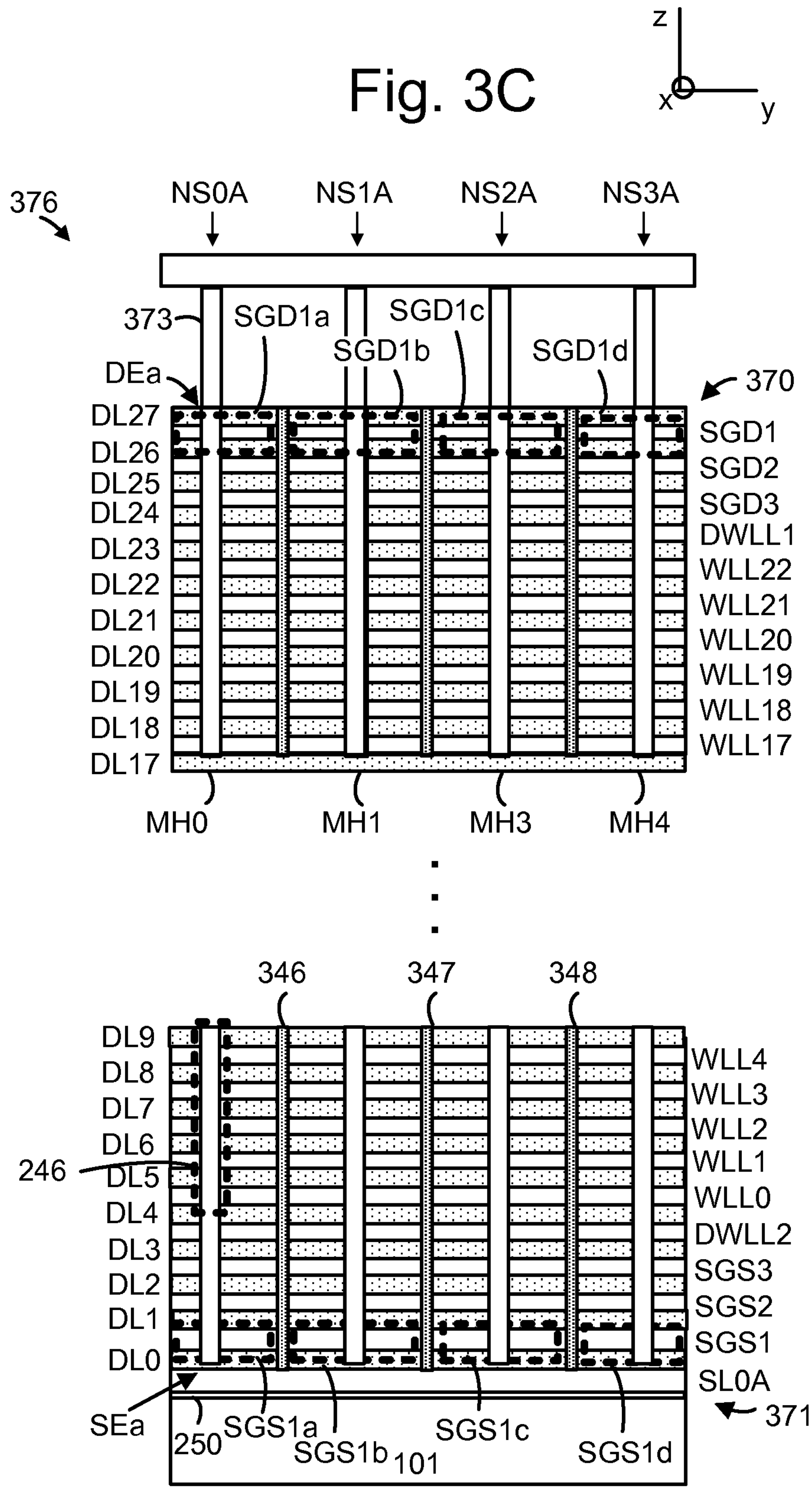

z
x
y
376
NS0A
NS1A
NS2A
NS3A
373
SGD1a
SGD1c
SGD1b
SGD1d
DEa
370
DL27
DL26
DL25
DL24
DL23
DL22
DL21
DL20
DL19
DL18
DL17
SGD1
SGD2
SGD3
DWLL1
WLL22
WLL21
WLL20
WLL19
WLL18
WLL17
MH0
MH1
MH3
MH4
346
347
348
DL9
DL8
DL7
DL6
246
DL5
DL4
DL3
DL2
DL1
DL0
WLL4
WLL3
WLL2
WLL1
WLL0
DWLL2
SGS3
SGS2
SGS1
SL0A
371
SEa
250
SGS1a
SGS1b
101
SGS1c
SGS1d

Fig. 4A

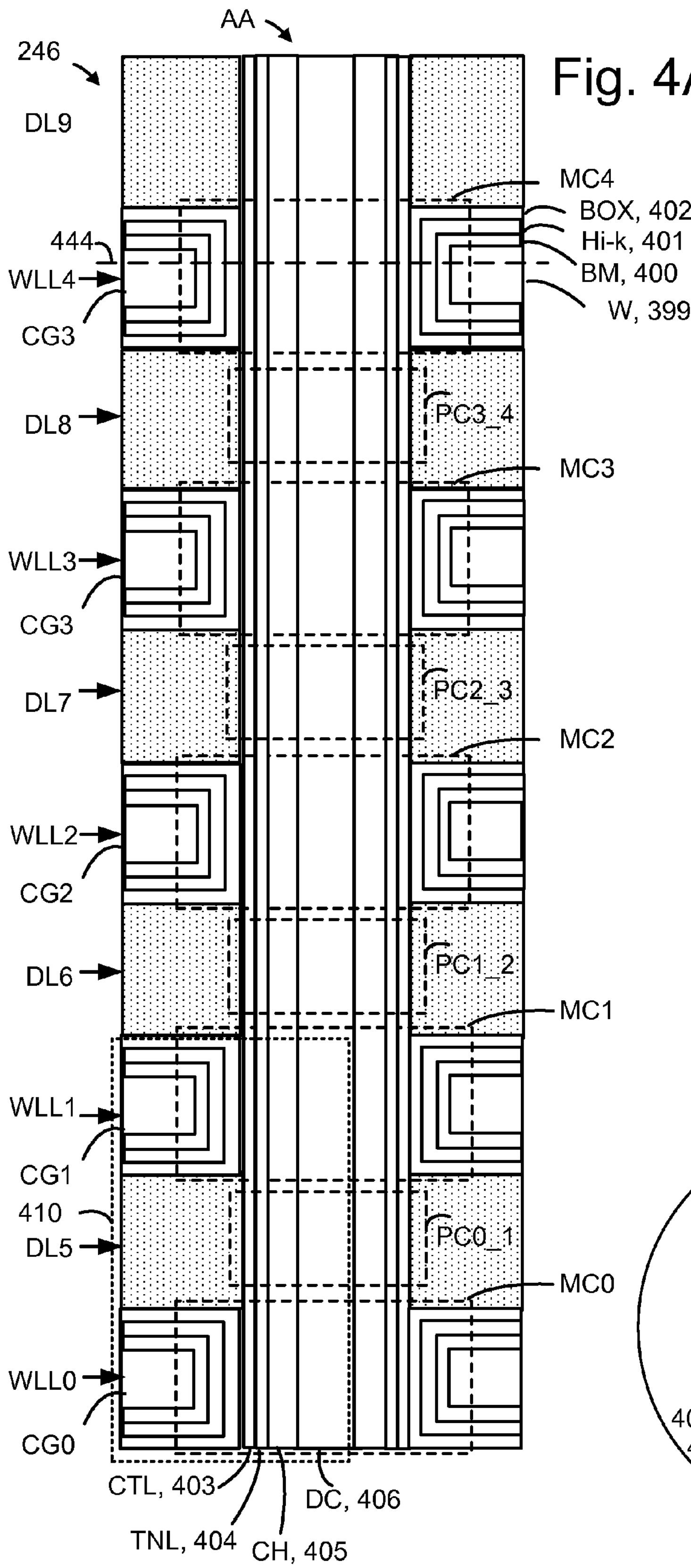

AA
246
DL9
MC4
BOX, 402
Hi-k, 401
BM, 400
W, 399
444
WLL4
CG3
DL8
PC3_4
MC3
WLL3
CG3
DL7
PC2_3
MC2
WLL2
CG2
DL6
PC1_2
MC1
WLL1
CG1
410
DL5
PC0_1
MC0
WLL0
CG0
CTL, 403
TNL, 404
CH, 405
DC, 406

Fig. 4B

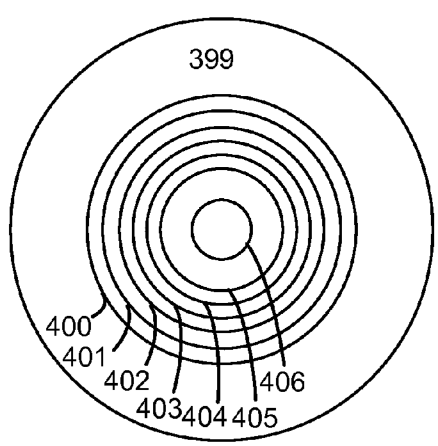

399
400
401
402
403
404
405
406

Fig. 4C

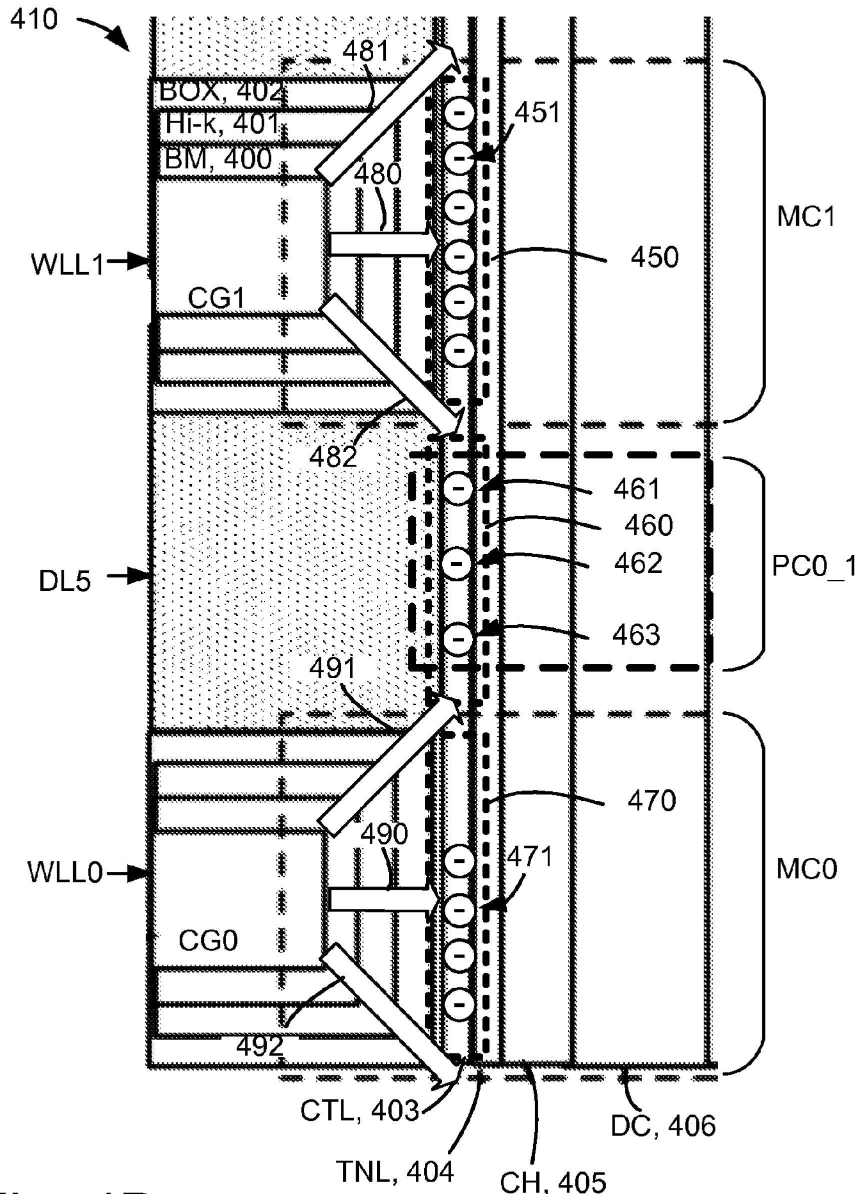


Fig. 4D

1st time			2nd time		
Vpgm(MC0)	Vpass(MC1)	Efield	Vpass(MC0)	Vpgm(MC1)	Efield
21 (A state)	9	30	7	21 (A state)	28
				22 (B state)	29
				23 (C state)	30
22 (B state)	8	30	7	21 (A state)	28
				22 (B state)	29
				23 (C state)	30
23 (C state)	7	30	7	21 (A state)	28
				22 (B state)	29
				23 (C state)	30

Fig. 5A

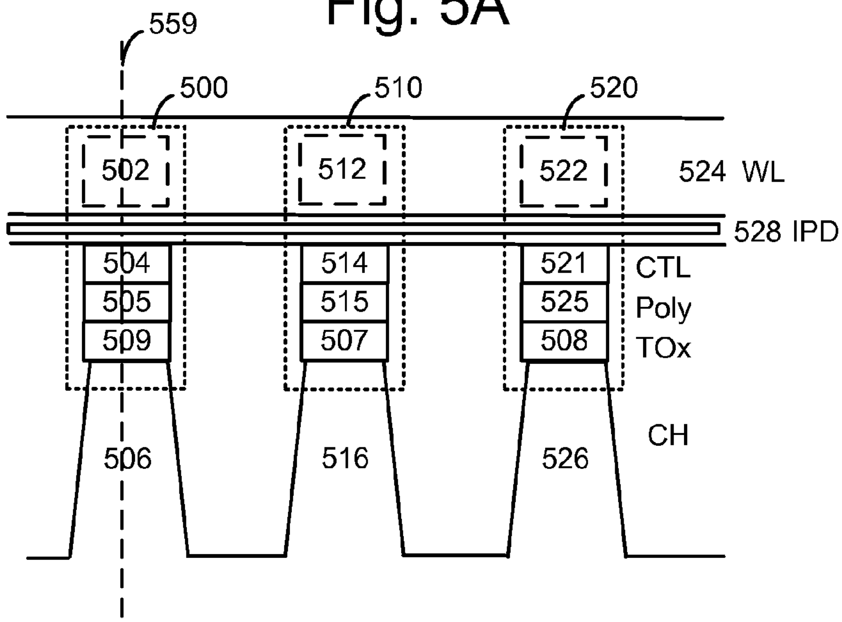
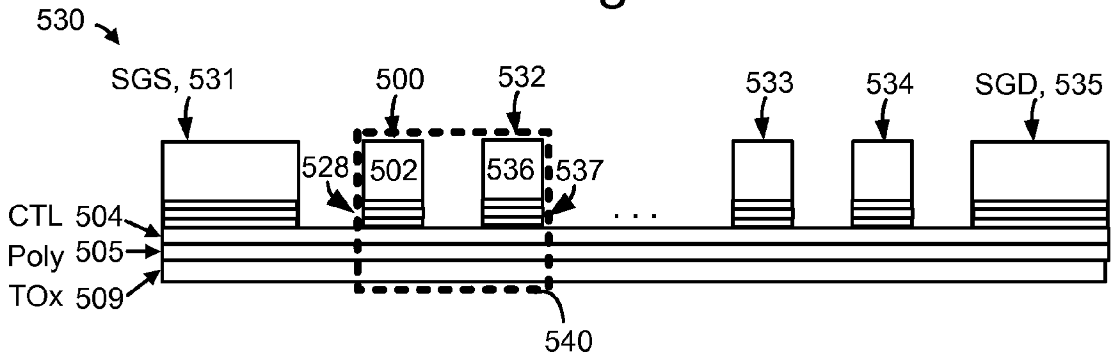


Fig. 5B



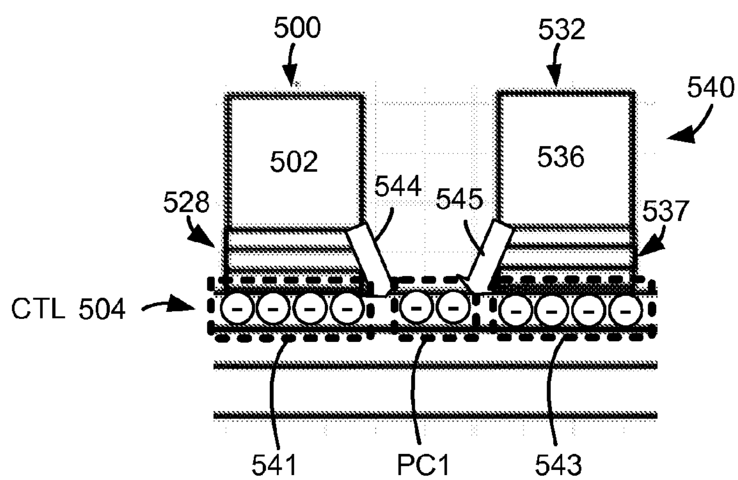


Fig. 5C

Fig. 6A

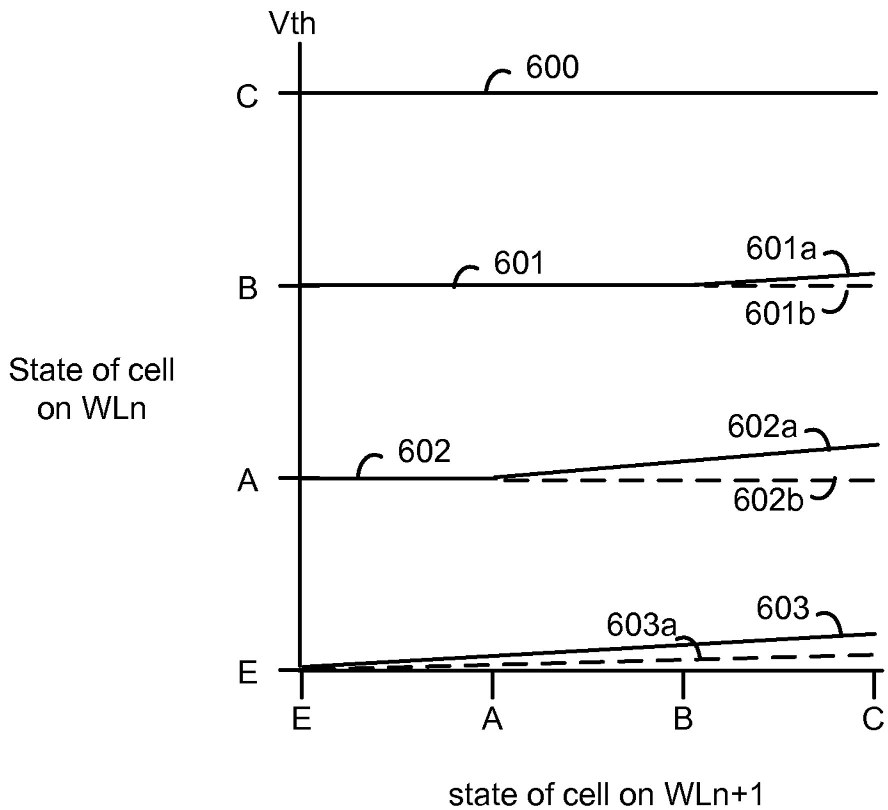

Vth
C
B
A
E
600
601
601a
601b
602
602a
602b
603
603a
State of cell
on WLn
E
A
B
C
state of cell on WLn+1

Fig. 6B

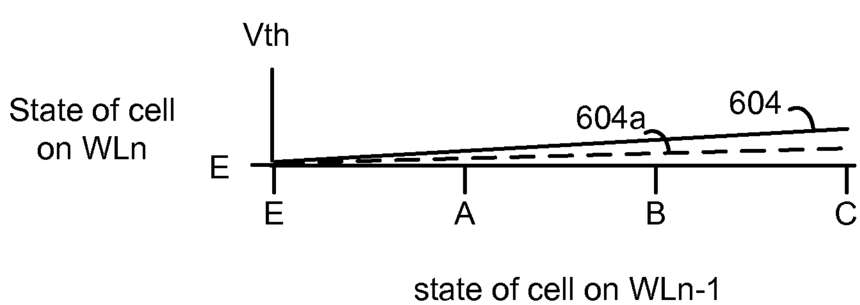
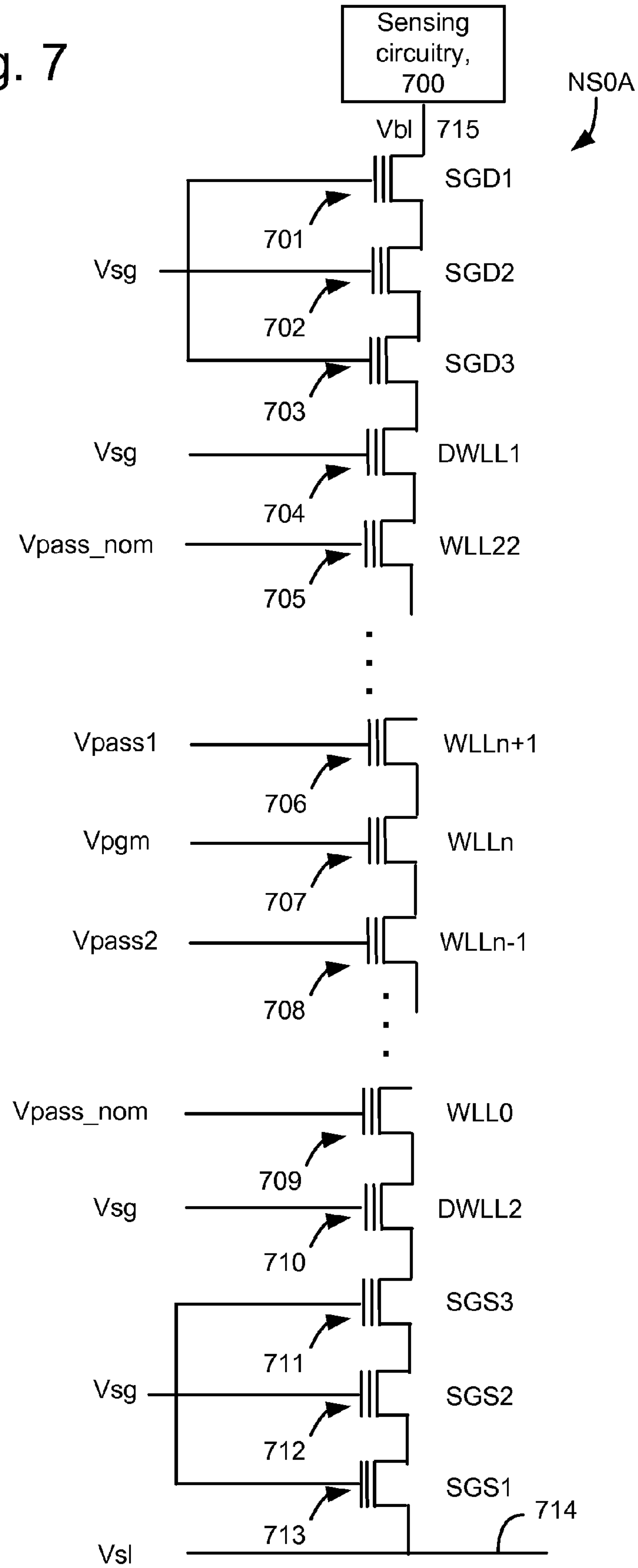

Vth
State of cell
on WLn
E
604a
604
E
A
B
C
state of cell on WLn-1

Fig. 7



Sensing circuitry, 700
NS0A
Vbl
715
SGD1
701
Vsg
SGD2
702
SGD3
703
Vsg
DWLL1
704
Vpass_nom
WLL22
705
Vpass1
WLLn+1
706
Vpgm
WLLn
707
Vpass2
WLLn-1
708
Vpass_nom
WLL0
709
Vsg
DWLL2
710
SGS3
711
Vsg
SGS2
712
SGS1
713
714
Vsl

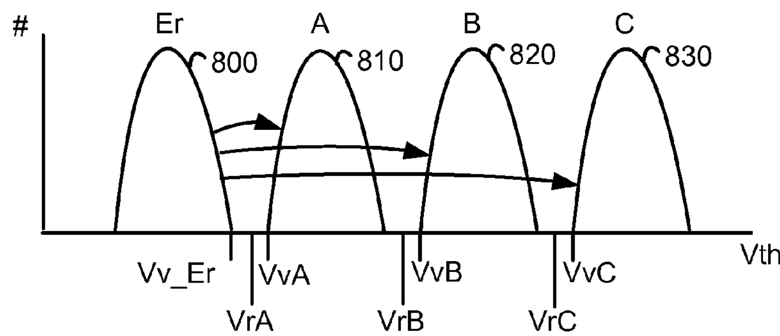


Fig. 8A

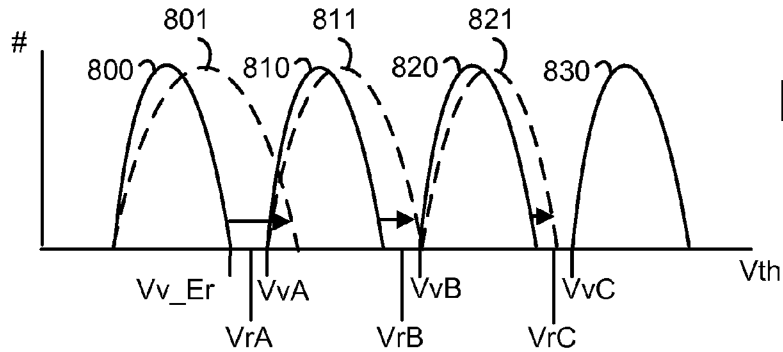


Fig. 8B

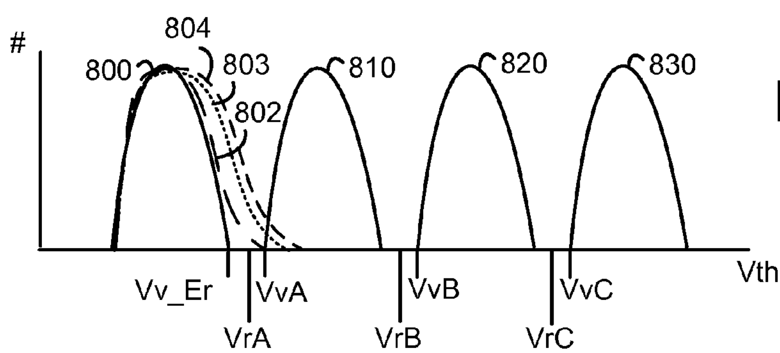


Fig. 8C

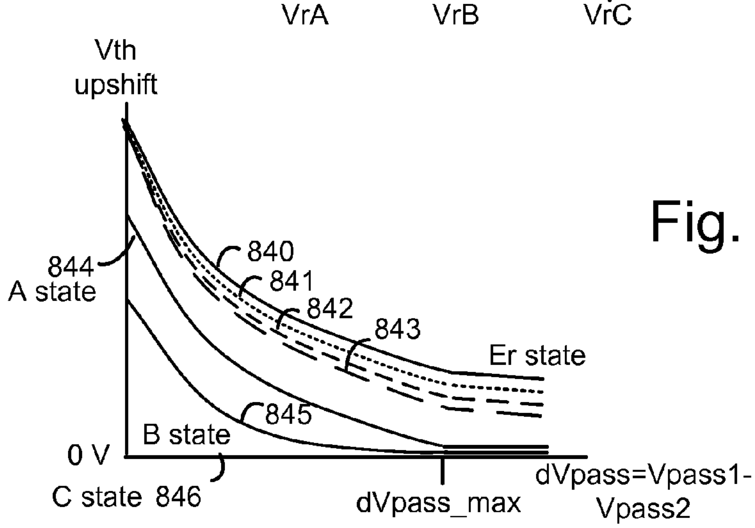


Fig. 8D

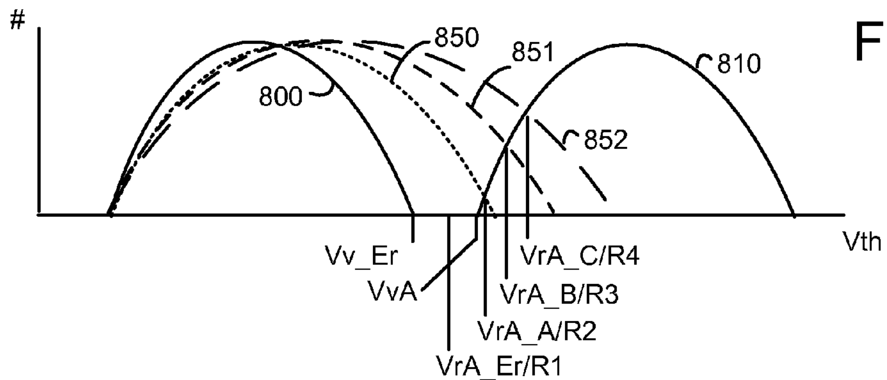


Fig. 8E

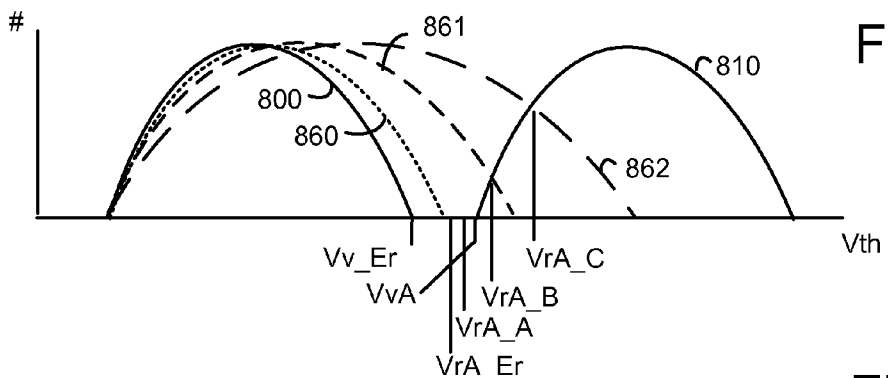
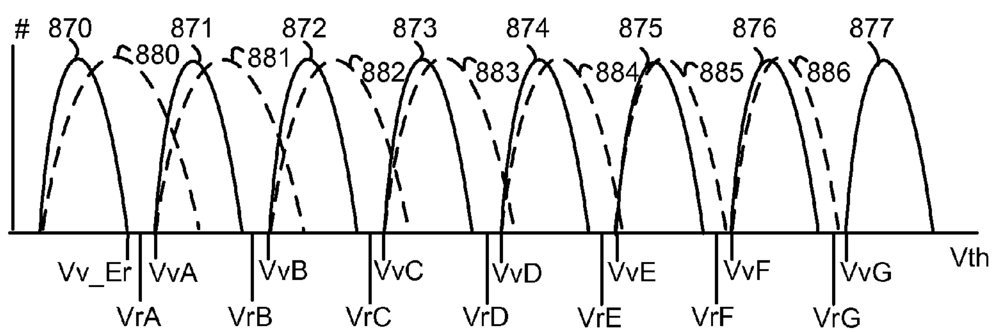


Fig. 8F

Fig. 8G



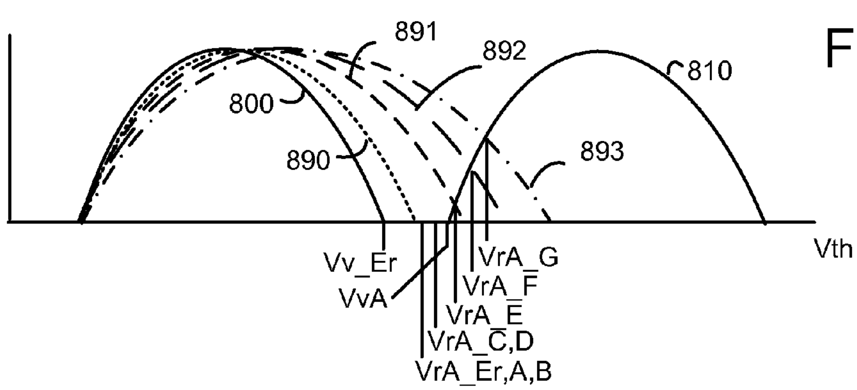


Fig. 8H

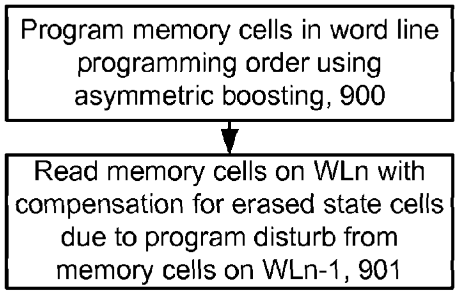

Program memory cells in word line programming order using asymmetric boosting, 900
Read memory cells on WLn with compensation for erased state cells due to program disturb from memory cells on WLn-1, 901

Fig. 9A

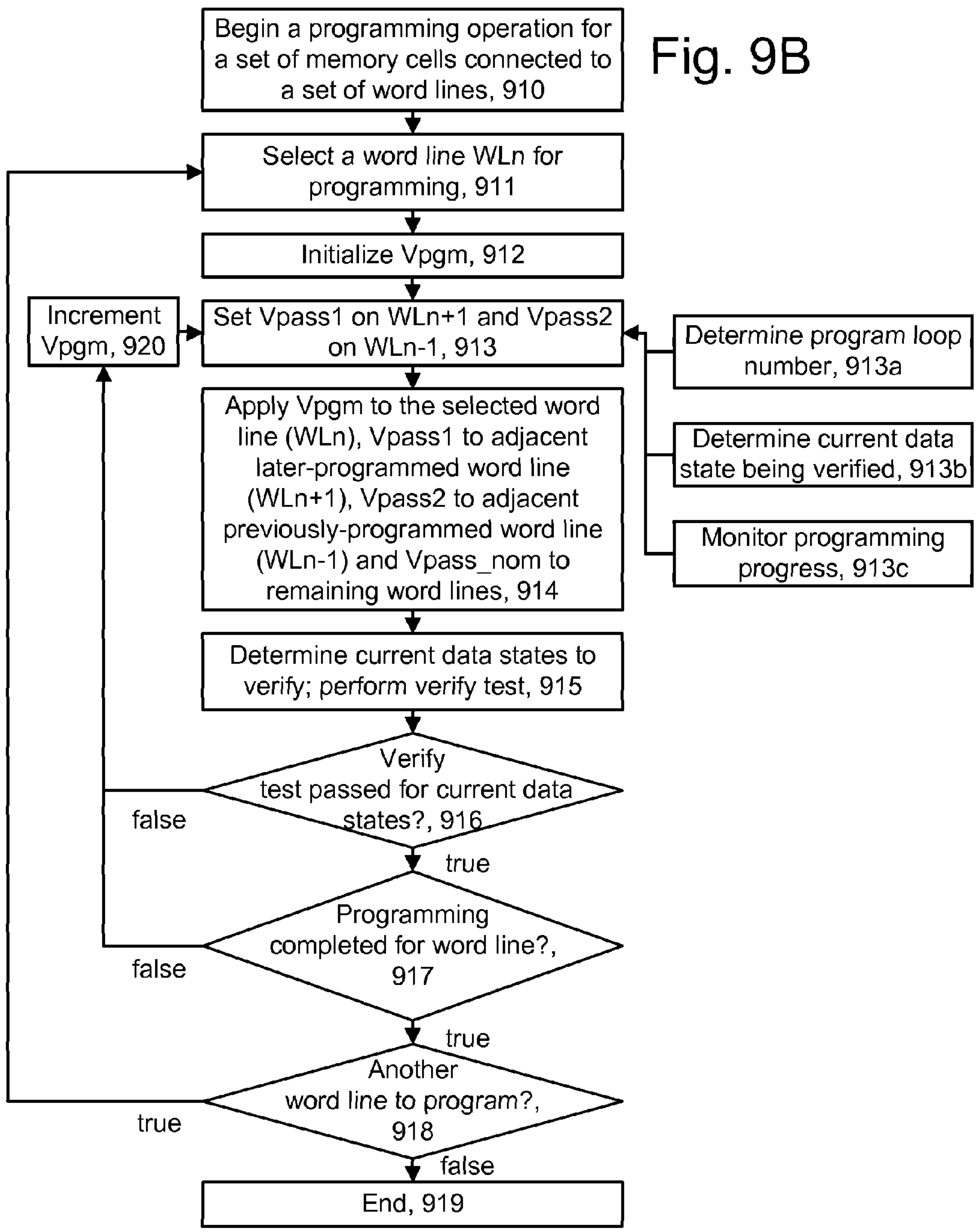

Begin a programming operation for a set of memory cells connected to a set of word lines, 910
Select a word line WLn for programming, 911
Initialize Vpgm, 912
Increment Vpgm, 920
Set Vpass1 on WLn+1 and Vpass2 on WLn-1, 913
Determine program loop number, 913a
Determine current data state being verified, 913b
Monitor programming progress, 913c
Apply Vpgm to the selected word line (WLn), Vpass1 to adjacent later-programmed word line (WLn+1), Vpass2 to adjacent previously-programmed word line (WLn-1) and Vpass_nom to remaining word lines, 914
Determine current data states to verify; perform verify test, 915
Verify test passed for current data states?, 916
false
true
Programming completed for word line?, 917
false
true
Another word line to program?, 918
true
false
End, 919

Fig. 9B

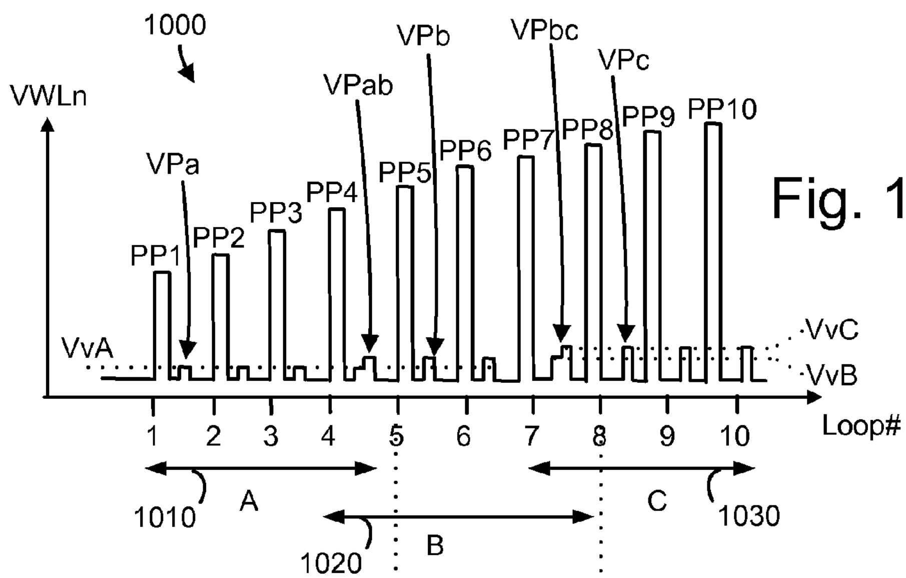


Fig. 10A

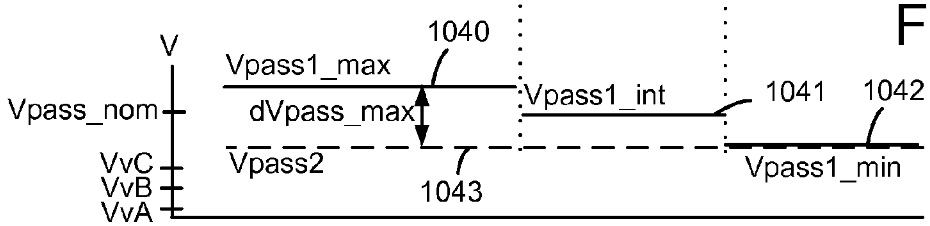


Fig. 10B

Fig. 11

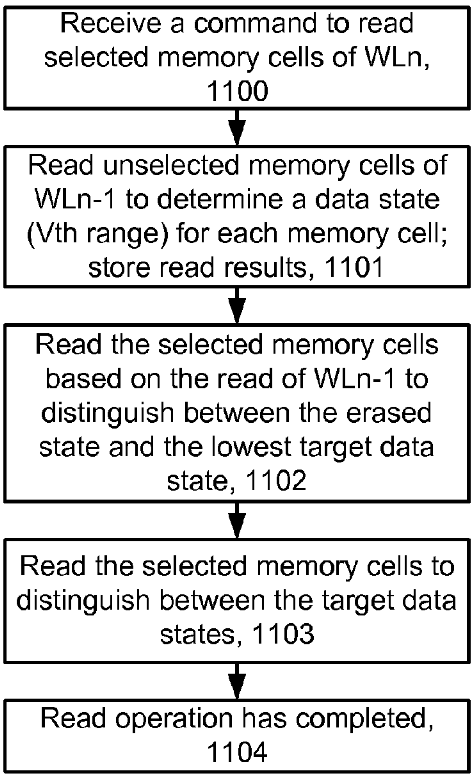


Fig. 12A

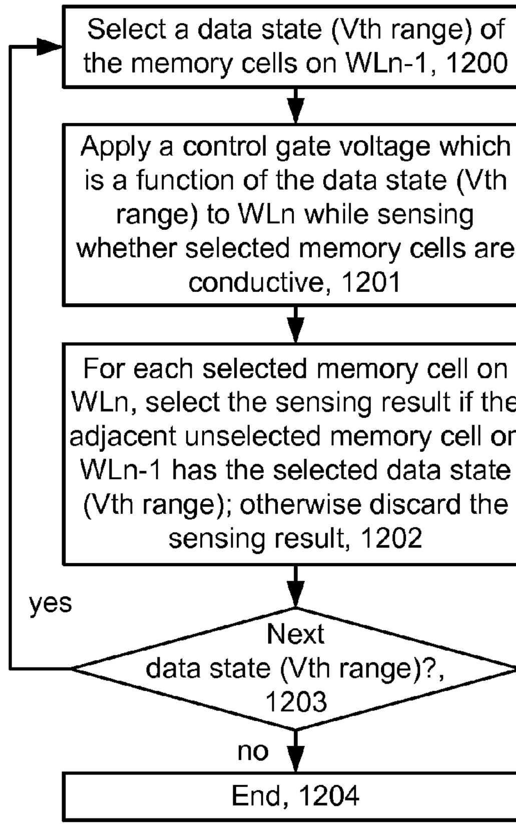


Fig. 12B

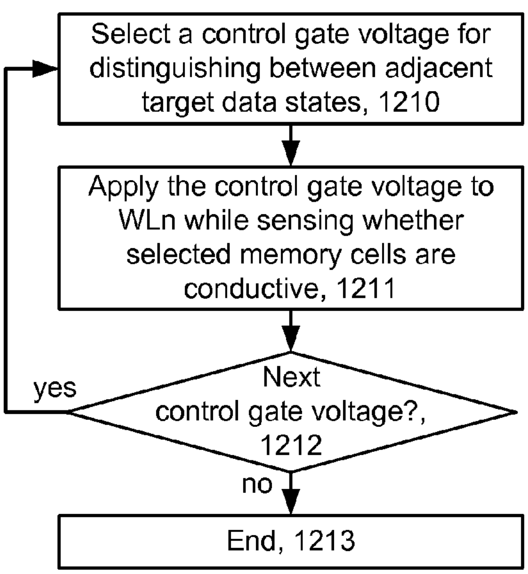


Fig. 13A

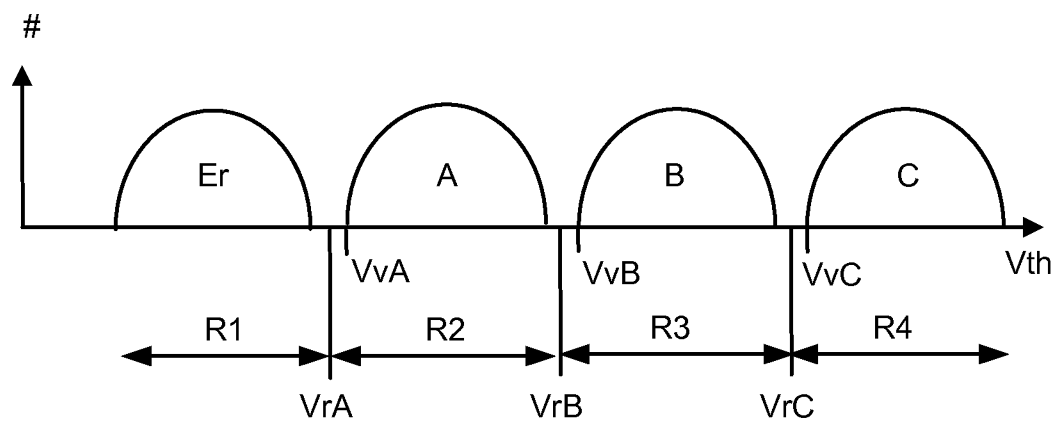


Fig. 13B

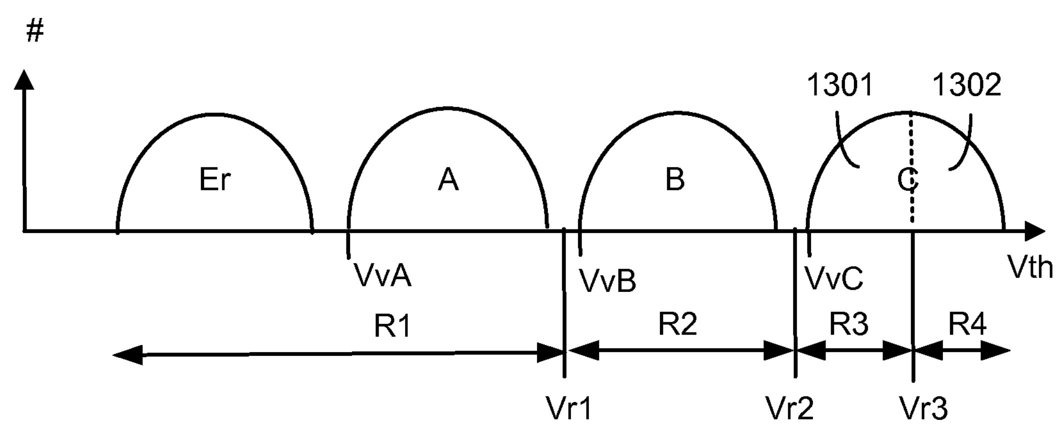


Fig. 14A

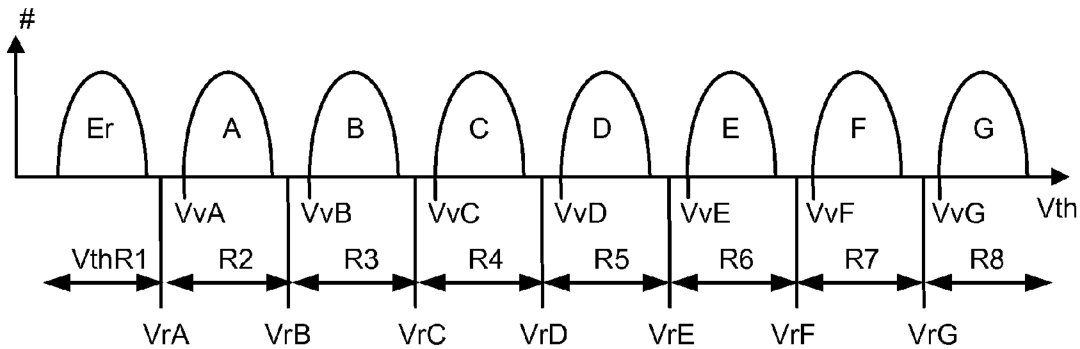


Fig. 14B

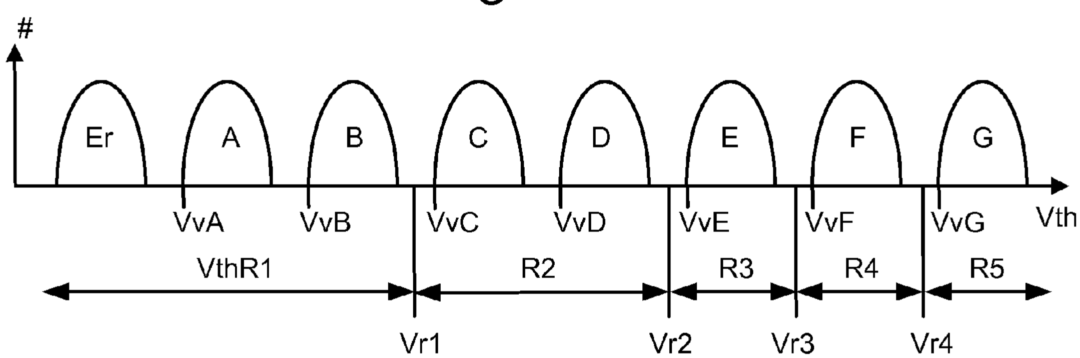


Fig. 15A

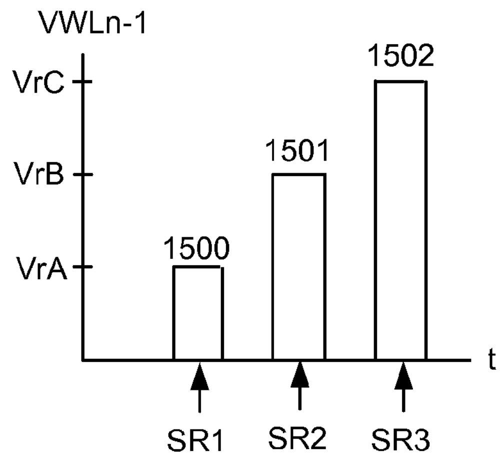


Fig. 15B

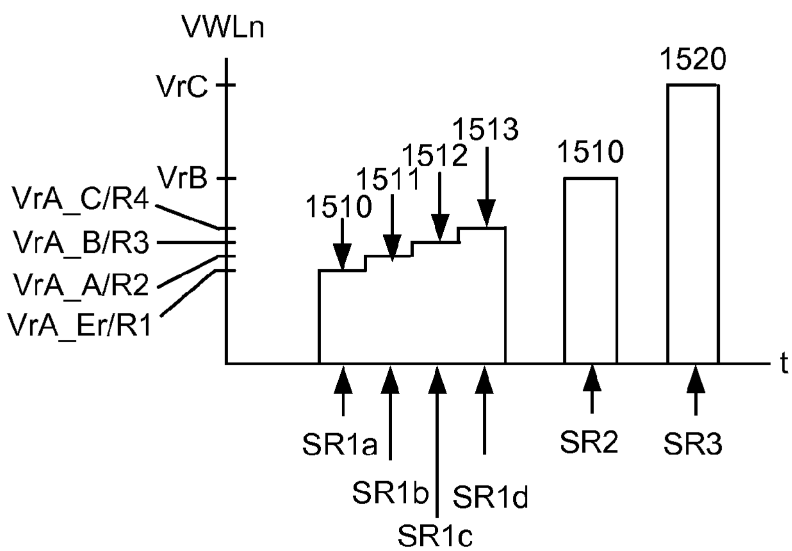


Fig. 15C

	WLn-1 state or Vth range:			
VWLn:	Er/R1	A/R2	B/R3	C/R4
VrA_Er/R1	HB	SB	SB	SB
VrA_A/R2	SB	HB	SB	SB
VrA_B/R3	SB	SB	HB	SB
VrA_C/R4	SB	SB	SB	HB

READ WITH LOOK-BACK COMBINED WITH PROGRAMMING WITH ASYMMETRIC BOOSTING IN MEMORY

BACKGROUND

The present technology relates to operation of memory devices.

A charge-trapping material can be used in memory devices to store a charge which represents a data state. The charge-trapping material can be arranged vertically in a three-dimensional (3D) stacked memory structure, or horizontally in a two-dimensional (2D) memory structure. One example of a 3D memory structure is the Bit Cost Scalable (BiCS) architecture which comprises a stack of alternating conductive and dielectric layers. A memory hole is formed in the stack and a NAND string is then formed by filling the memory hole with materials including a charge-trapping layer. A straight NAND string extends in one memory hole, while a pipe- or U-shaped NAND string (P-BiCS) includes a pair of vertical columns of memory cells which extend in two memory holes and which are joined by a bottom back gate. Control gates of the memory cells are provided by the conductive layers.

However, various challenges are presented in operating such memory devices.

BRIEF DESCRIPTION OF THE DRAWINGS

Like-numbered elements refer to common components in the different figures.

FIG. **1**A is a perspective view of a 3D stacked non-volatile memory device.

FIG. **1**B is a functional block diagram of a memory device such as the 3D stacked non-volatile memory device **100** of FIG. **1**A or a 2D memory device such as depicted in FIG. **5**A to **5**C.

FIG. **2**A depicts a top view of example word line layers **202** and **204** in a U-shaped NAND embodiment, as an example implementation of BLK**0** in FIG. **1**A.

FIG. **2**B depicts a top view of example select gate layer portions, consistent with FIG. **2**A.

FIG. **2**C depicts an embodiment of a stack **231** showing a cross-sectional view of the portion **209** of FIG. **2**A, along line **220**.

FIG. **3**A depicts a top view of an example word line layer **304** of the block BLK**0** of FIG. **1**A, in a straight NAND string embodiment.

FIG. **3**B depicts a top view of an example SGD layer **362**, consistent with FIG. **3**A.

FIG. **3**C depicts an embodiment of a stack **376** showing a cross-sectional view of the portion **307** of FIG. **3**A, along line **305**.

FIG. **4**A depicts a view of the region **246** of FIG. **3**C, showing memory cells MC**1** to MC**5**.

FIG. **4**B depicts a cross-section view of the region **246** of FIG. **4**A along line **444**.

FIG. **4**C depicts an expanded view of a portion **410** of the NAND string of FIG. **4**A, showing the formation of a parasitic cell PC**0_1** between the memory cells MC**0** and MC**1**.

FIG. **4**D depicts a table showing a magnitude of an electric field (Efield) which is seen by the parasitic cell PC**0_1** of FIG. **4**C with various combinations of pass voltages and data states on MC**0** and MC**1**.

FIG. **5**A depicts a cross-sectional view in a word line direction of memory cells comprising a flat control gate and charge-trapping regions as a 2D example of memory cells in the memory structure **126** of FIG. **1**B.

FIG. **5**B depicts a cross sectional view along line **559** in FIG. **5**A, showing a NAND string **530** having a flat control gate and a charge-trapping layer.

FIG. **5**C depicts an expanded view of a portion **540** of the NAND string of FIG. **5**B, showing the formation of a parasitic cell PC**1**.

FIG. **6**A depicts a plot showing an increase in the Vth of a memory cell on a selected word line WLn due to program disturb, as a function of a data state of a memory cell on a later-programmed word line, WLn+1.

FIG. **6**B depicts a plot showing an increase in the Vth of an erased state memory cell on a selected word line WLn as a function of a data state of a memory cell on a previously-programmed word line, WLn−1.

FIG. **7** depicts a circuit diagram of a NAND string consistent with the memory devices of FIGS. **2**C, **3**C and **5**B.

FIG. **8**A depicts threshold voltage (Vth) distributions of memory cells in an example embodiment in which there is an erased state and three target data states.

FIG. **8**B depicts widening of the Vth distributions of FIG. **8**A due to program disturb, in a symmetric boosting case.

FIG. **8**C depicts a reduction in the widening of the Vth distributions of FIG. **8**A due to program disturb, in an asymmetric boosting case.

FIG. **8**D depicts a plot showing a reduction in the Vth upshift for different data states on WLn as a function of dVpass=Vpass**1**−Vpass**2**.

FIG. **8**E depicts an example in which different values of a read voltage are used to account for different levels of disturb to memory cells in the erased state on WLn as a function of four possible data states of memory cells on WLn−1.

FIG. **8**F depicts another example in which different values of a read voltage are used to account for different levels of disturb to memory cells in the erased state on WLn as a function of four possible data states of memory cells on WLn−1.

FIG. **8**G depicts Vth distributions of memory cells in an example embodiment in which there is an erased state and seven target data states.

FIG. **8**H depicts an example in which different values of a read voltage are used to account for different levels of disturb to memory cells in the erased state on WLn as a function of eight possible data states of memory cells on WLn−1.

FIG. **9**A depicts a process for programming and then reading memory cells in a memory device.

FIG. **9**B depicts an example programming operation for memory cells of a selected word line WLn which minimizes program disturb by optimizing pass voltages, consistent with step **900** of FIG. **9**A.

FIG. **10**A depicts a plurality of program-verify iterations consistent with the programming operation of FIG. **9**B.

FIG. **10**B depicts one example of pass voltages consistent with FIG. **10**A in which a decreasing level of Vpass**1** is used on WLn+1 and a fixed level of Vpass**2** is used on WLn−1.

FIG. **11** depicts an example of a reading operation consistent with step **901** of FIG. **9**A.

FIG. **12**A depicts a reading process consistent with step **1102** of FIG. **11**.

FIG. **12**B depicts a reading process consistent with step **1103** of FIG. **11**.

FIG. **13**A depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with one Vth range per data state and four total data states, consistent with step **1101** of FIG. **11**.

FIG. **13**B depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with two Vth ranges for the highest data state (C), one Vth range for the next data state

(B), and one Vth range for the lowest two data states (Er and A), and four total data states, consistent with step **1101** of FIG. **11**.

FIG. **14**A depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with one Vth range per data state and eight total data states, consistent with step **1101** of FIG. **11**.

FIG. **14**B depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with one Vth range for each of the three highest data states (E, F and G), one Vth range for the next two highest data states (C and D), and one Vth range for the lowest three data states (Er, A and B), and eight total data states, consistent with step **1101** of FIG. **11**.

FIG. **15**A depicts control gate voltages applied to an unselected word line, consistent with step **1101** of FIG. **11**.

FIG. **15**B depicts control gate voltages applied to a selected word line, consistent with steps **1102** and **1103** of FIG. **11**.

FIG. **15**C depicts a table showing hard bits and soft bits consistent with FIG. **15**B.

DETAILED DESCRIPTION

Techniques are provided for reducing program disturb during programming and reading of memory cells which are formed along a continuous charge-trapping region in a memory device.

In a charge-trapping memory device, a continuous charge-trapping region may extend throughout a NAND string, for instance, where different portions of the charge-trapping region are associated with different respective memory cells which are connected to different respective word lines. Each memory cell has a control gate which receives a program voltage (Vpgm) via respective word line when it is programmed. For each programmed memory cell, the program voltage causes charges (electrons) to move from a channel region to a respective portion of the charge-trapping region, where the charges are stored to represent a programmed data state of the memory cell. Some cells are not programmed and remain in an erased state. While a selected memory cell receives the program voltage, other, unselected memory cells receive a pass voltage which is high enough to provide the unselected memory cells in a conductive state but not high enough to program them. Moreover, typically the memory cells are programmed in a word line programming order, one word line at a time, starting at one end of a NAND string and continuing to the other end of the NAND string.

However, during a programming operation, portions of the charge-trapping region which are between the memory cells can be inadvertently programmed. In particular, the highest program voltage on a word line and the pass voltage on an adjacent word line work in combination to produce a fringing electric field which can draw electrons into a portion of the charge-trapping region which is between the word lines. The highest program voltages are used to complete the programming of the highest data states, e.g., the C state in a four level memory device comprising Er, A, B and C states, or the F and G states in an eight level memory device comprising Er, A, B, C, D, E, F and G states.

This portion of the charge-trapping region which is between the word lines acts as a parasitic cell. When the memory cells are subsequently read, they may experience an upshift in their threshold voltage due to the parasitic cells. The memory cells in the erased state and the lowest target data state (e.g., A state) are subject to the largest Vth upshift. Moreover, this problem is expected to become worse as the distance between word lines decreases.

Techniques provided herein address the above-mentioned issues. One aspect involves the programming operation. The techniques involve optimizing the pass voltages during a programming operation so that, when a selected nth word line (WLn) is programmed, the pass voltage is higher on a later-programmed word line (e.g., WLn+1) than a previously-programmed word line (e.g., WLn−1) during an initial part of the programming operation. This is referred to as asymmetric boosting. Subsequently, the pass voltage on WLn+1 can be gradually lowered. This approach notes that a parasitic cell is subject to a fringing field in two situations. For example, consider a parasitic cell (e.g., PC**0**_**1** in FIG. **4**C) which is between a memory cell MC**0** on a word line WLL**0** and a memory cell MC**1** on a word line WLL**1**.

In a first situation, when MC**0** is completing programming, MC**0** receives a final Vpgm, Vpgm(MC**0**), and MC**1** receives a pass voltage, Vpass**1**. PC**0**_**1** therefore will have a Vth which is a function of Vpgm(MC**0**)+Vpass**1**. In a second situation, when MC**1** is completing programming, MC**1** receives a final Vpgm, Vpgm(MC**1**), and MC**0** receives a pass voltage Vpass**2**. PC**0**_**1** will have a Vth which is a function of Vpgm(MC**1**)+Vpass**2**, if Vpgm(MC**1**)+Vpass**2**>Vpgm(MC**0**)+Vpass**1**. This is true because the parasitic cell is not programmed further unless it is exposed to a stronger fringing field than what it has previously experienced.

To avoid further programming of PC**0**_**1** when MC**1** is programmed, we have: Vpass**1**−Vpass**2**>=Vpgm(MC**1**)−Vpgm(MC**0**). Thus, a difference between the pass voltages can be set based on a highest possible difference between the highest final Vpgm on the two memory cells. This difference typically correlates to the difference between the verify levels of the highest and lowest programmed data states. Further, Vpass should not be too high or it can begin to program a memory cell and Vpass should not be too low or it will provide an inadequate channel boosting potential.

When a given memory cell is read, its Vth is affected by the parasitic cell which is after the given memory cell in the programming order, because this parasitic cell may be further programmed after the given memory cell is programmed. In contrast, the parasitic cell which is before the given memory cell in the programming order generally does not affect the Vth of the given memory cell when it is read because any effect of this parasitic cell can be compensated for in the programming of the given memory cell. However, if the given memory cell is in the erased state, its Vth is affected by the parasitic cells which are before and after the given memory cell in the programming order. To minimize the programming of the parasitic cells adjacent to an erased state cell, Vpass**1** can be lowered as the programming operation proceeds, so that it is at its lowest point when the highest Vpgm is applied. Specifically, Vpass**1** can step down so that on or before the final program pulse of the programming operation, it reaches a minimum level which is the same as, or slightly higher than (e.g., up to 1 V higher) Vpass**2**. That is, Vpass**1** is no lower than the second pass voltage during a final program voltage of the program voltages. Vpass**2** can remain at a fixed level throughout the programming operation, in one approach.

Providing Vpass**1**>Vpass**2** initially, and lowering Vpass**1** (and the difference dVpass=Vpass**1**−Vpass**2**) as the programming operation proceeds, optimally reduces the widening of the erased state cells and the target data state cells.

Another aspect involves the read operation, which is optimized in combination with the asymmetric boosting of the programming operation. The read operation for a memory cell on a word line WLn involves a plurality of sense operations to distinguish between the Er and A states, a single sense operation to distinguish between the A and B states, and another single sense operation to distinguish between the B and C states. An optimal result from one of the plurality of sense operations is selected based on the data state of an adjacent memory cell on WLn−1. Based on the optimal result for distinguishing between the Er and A states and the single sense operations which distinguish between the A and B states, and the B and C states, the data state of the memory cell on WLn can be determined. Optionally, the optimal result from one of the plurality of sense operations can be selected based on a Vth range of the adjacent memory cell on WLn−1. A Vth range can encompass part of a data state, exactly one data state, or multiple data states, for instance.

The following discussion provides details of the construction of example memory devices and of related techniques which address the above and other issues.

FIG. **1**A is a perspective view of a 3D stacked non-volatile memory device. The memory device **100** includes a substrate **101**. On the substrate are example blocks BLK**0** and BLK**1** of memory cells and a peripheral area **104** with circuitry for use by the blocks. The substrate **101** can also carry circuitry under the blocks, along with one or more lower metal layers which are patterned in conductive paths to carry signals of the circuitry. The blocks are formed in an intermediate region **102** of the memory device. In an upper region **103** of the memory device, one or more upper metal layers are patterned in conductive paths to carry signals of the circuitry. Each block comprises a stacked area of memory cells, where alternating levels of the stack represent word lines. In one possible approach, each block has opposing tiered sides from which vertical contacts extend upward to an upper metal layer to form connections to conductive paths. While two blocks are depicted as an example, additional blocks can be used, extending in the x- and/or y-directions.

In one possible approach, the length of the plane, in the x-direction, represents a direction in which signal paths to word lines extend in the one or more upper metal layers (a word line or SGD line direction), and the width of the plane, in the y-direction, represents a direction in which signal paths to bit lines extend in the one or more upper metal layers (a bit line direction). The z-direction represents a height of the memory device.

FIG. **1**B is a functional block diagram of a memory device such as the 3D stacked non-volatile memory device **100** of FIG. **1**A or a 2D memory device such as depicted in FIG. **5**A to **5**C. The memory device **100** may include one or more memory die **108**. The memory die **108** includes a memory structure **126** of memory cells, such as an array of cells, control circuitry **110**, and read/write circuits **128**. In a 3D configuration, the memory array can include the blocks BLK**0** and BLK**1** of FIG. **1**A. The memory structure **126** is addressable by word lines via a row decoder **124** and by bit lines via a column decoder **132**. The read/write circuits **128** include multiple sense blocks **130** (sensing circuitry) and allow a page of memory cells to be read or programmed in parallel. Typically a controller **122** is included in the same memory device **100** (e.g., a removable storage card) as the one or more memory die **108**. Commands and data are transferred between the host and controller **122** via lines **120** and between the controller and the one or more memory die **108** via lines **118**.

The memory structure may comprise one or more array of memory cells including a 3D array. The memory structure may comprise a monolithic three dimensional memory array in which multiple memory levels are formed above (and not in) a single substrate, such as a wafer, with no intervening substrates. The memory structure may comprise any type of non-volatile memory that is monolithically formed in one or more physical levels of arrays of memory cells having an active area disposed above a silicon substrate. The memory structure may be in a non-volatile memory device having circuitry associated with the operation of the memory cells, whether the associated circuitry is above or within the substrate.

The control circuitry **110** cooperates with the read/write circuits **128** to perform memory operations on the memory structure **126**, and includes a state machine **112**, an on-chip address decoder **114**, and a power control module **116**. The state machine **112** provides chip-level control of memory operations. A storage region **113** may be provided for pass voltages and read voltages in the memory device. This can include information regarding which level of a pass voltage to apply during a specific program loop or read operation and which read levels to use in a read process.

The on-chip address decoder **114** provides an address interface between that used by the host or a memory controller to the hardware address used by the decoders **124** and **132**. The power control module **116** controls the power and voltages supplied to the word lines and bit lines during memory operations. It can includes drivers for word line layers (WLLs) in a 3D configuration, SGS and SGD transistors and source lines. The sense blocks **130** can include bit line drivers, in one approach. An SGS transistor is a select gate transistor at a source end of a NAND string, and an SGD transistor is a select gate transistor at a drain end of a NAND string.

In some implementations, some of the components can be combined. In various designs, one or more of the components (alone or in combination), other than memory structure **126**, can be thought of as at least one control circuit which is configured to perform the actions described herein. For example, a control circuit may include any one of, or a combination of, control circuitry **110**, state machine **112**, decoders **114**/**132**, power control module **116**, sense blocks **130**, read/write circuits **128**, and controller **122**, and so forth.

Other types of non-volatile memory in addition to NAND flash memory can also be used.

Semiconductor memory devices include volatile memory devices, such as dynamic random access memory ("DRAM") or static random access memory ("SRAM") devices, non-volatile memory devices, such as resistive random access memory ("ReRAM"), electrically erasable programmable read only memory ("EEPROM"), flash memory (which can also be considered a subset of EEPROM), ferroelectric random access memory ("FRAM"), and magnetoresistive random access memory ("MRAM"), and other semiconductor elements capable of storing information. Each type of memory device may have different configurations. For example, flash memory devices may be configured in a NAND or a NOR configuration.

The memory devices can be formed from passive and/or active elements, in any combinations. By way of non-limiting example, passive semiconductor memory elements include ReRAM device elements, which in some embodiments include a resistivity switching storage element, such as an anti-fuse or phase change material, and optionally a steering element, such as a diode or transistor. Further by way of non-limiting example, active semiconductor memory elements include EEPROM and flash memory device elements, which in some embodiments include elements containing a charge storage region, such as a floating gate, conductive nanoparticles, or a charge storage dielectric material.

Multiple memory elements may be configured so that they are connected in series or so that each element is individually accessible. By way of non-limiting example, flash memory devices in a NAND configuration (NAND memory) typically contain memory elements connected in series. A NAND memory array may be configured so that the array is composed of multiple strings of memory in which a string is composed of multiple memory elements sharing a single bit line and accessed as a group. Alternatively, memory elements may be configured so that each element is individually accessible, e.g., a NOR memory array. NAND and NOR memory configurations are exemplary, and memory elements may be otherwise configured.

The semiconductor memory elements located within and/or over a substrate may be arranged in two or three dimensions, such as a two dimensional memory structure or a three dimensional memory structure.

In a two dimensional memory structure, the semiconductor memory elements are arranged in a single plane or a single memory device level. Typically, in a two dimensional memory structure, memory elements are arranged in a plane (e.g., in an x-y direction plane) which extends substantially parallel to a major surface of a substrate that supports the memory elements. The substrate may be a wafer over or in which the layer of the memory elements are formed or it may be a carrier substrate which is attached to the memory elements after they are formed. As a non-limiting example, the substrate may include a semiconductor such as silicon.

The memory elements may be arranged in the single memory device level in an ordered array, such as in a plurality of rows and/or columns. However, the memory elements may be arrayed in non-regular or non-orthogonal configurations. The memory elements may each have two or more electrodes or contact lines, such as bit lines and word lines.

A three dimensional memory array is arranged so that memory elements occupy multiple planes or multiple memory device levels, thereby forming a structure in three dimensions (i.e., in the x, y and z directions, where the z direction is substantially perpendicular and the x and y directions are substantially parallel to the major surface of the substrate).

As a non-limiting example, a three dimensional memory structure may be vertically arranged as a stack of multiple two dimensional memory device levels. As another non-limiting example, a three dimensional memory array may be arranged as multiple vertical columns (e.g., columns extending substantially perpendicular to the major surface of the substrate, i.e., in the y direction) with each column having multiple memory elements. The columns may be arranged in a two dimensional configuration, e.g., in an x-y plane, resulting in a three dimensional arrangement of memory elements with elements on multiple vertically stacked memory planes. Other configurations of memory elements in three dimensions can also constitute a three dimensional memory array.

By way of non-limiting example, in a three dimensional NAND memory array, the memory elements may be coupled together to form a NAND string within a single horizontal (e.g., x-y) memory device level. Alternatively, the memory elements may be coupled together to form a vertical NAND string that traverses across multiple horizontal memory device levels. Other three dimensional configurations can be envisioned wherein some NAND strings contain memory elements in a single memory level while other strings contain memory elements which span through multiple memory levels. Three dimensional memory arrays may also be designed in a NOR configuration and in a ReRAM configuration.

Typically, in a monolithic three dimensional memory array, one or more memory device levels are formed above a single substrate. Optionally, the monolithic three dimensional memory array may also have one or more memory layers at least partially within the single substrate. As a non-limiting example, the substrate may include a semiconductor such as silicon. In a monolithic three dimensional array, the layers constituting each memory device level of the array are typically formed on the layers of the underlying memory device levels of the array. However, layers of adjacent memory device levels of a monolithic three dimensional memory array may be shared or have intervening layers between memory device levels.

Then again, two dimensional arrays may be formed separately and then packaged together to form a non-monolithic memory device having multiple layers of memory. For example, non-monolithic stacked memories can be constructed by forming memory levels on separate substrates and then stacking the memory levels atop each other. The substrates may be thinned or removed from the memory device levels before stacking, but as the memory device levels are initially formed over separate substrates, the resulting memory arrays are not monolithic three dimensional memory arrays. Further, multiple two dimensional memory arrays or three dimensional memory arrays (monolithic or non-monolithic) may be formed on separate chips and then packaged together to form a stacked-chip memory device.

Associated circuitry is typically required for operation of the memory elements and for communication with the memory elements. As non-limiting examples, memory devices may have circuitry used for controlling and driving memory elements to accomplish functions such as programming and reading. This associated circuitry may be on the same substrate as the memory elements and/or on a separate substrate. For example, a controller for memory read-write operations may be located on a separate controller chip and/or on the same substrate as the memory elements.

One of skill in the art will recognize that this technology is not limited to the two dimensional and three dimensional exemplary structures described but covers all relevant memory structures within the spirit and scope of the technology as described herein and as understood by one of skill in the art.

FIG. **2**A depicts a top view of example word line layers **202** and **204** in a U-shaped NAND embodiment, as an example implementation of BLK**0** in FIG. **1**A. In a 3D stacked memory device, memory cells are formed along memory holes which extend through alternating conductive and dielectric layers in a stack. The memory cells are typically arranged in NAND strings. Each conductive layer can include one or more word line layers. A word line layer is an example of a word line.

The view is of a representative layer among the multiple WLLs in a stack. Referring also to FIG. **2**C, the stack includes alternating dielectric and conductive layers. The dielectric layers include DL**0** to DL**25** and may be made of SiO_2, for instance. The conductive layers include a back gate layer (BGL), data-storing word line layers WLL**0** to WLL**19**, dummy (non-data-storing) word line layers DWLLa and DWLLb, and select gate layers SGL**1**, SGL**2** and SGL**3**. The word line layers are conductive paths to control gates of the memory cells at the layer. Moreover, each select gate layer may comprises conductive lines to select gate transistors (e.g., SGD and/or SGS transistors).

The word line layers of FIG. **2**A may represent any one of the word line layers in FIG. **2**C. These conductive layers may include doped polysilicon, metal such as tungsten or metal silicide, for instance. An example voltage of 5-10 V may be applied to the back gate to maintain a conductive state which connects the drain- and source-side columns.

For each block, each conductive layer may be divided into two word line layers **202** and **204** which are insulated from one another by a slit **206**. The slit is formed by etching a void which extends vertically in the stack, typically from an etch stop layer at the bottom to at least a top layer of the stack, then filling the slit with insulation. This is an example of the type of etching which can result in the accumulation of charges in the top conductive layer of the stack. The slit **206** is a single continuous slit which extends in a zig-zag pattern in the block. This approach can provide greater flexibility in controlling the memory cells since the WLLs can be driven independently.

Each block includes memory holes or pillars which extend vertically in the stack, and comprise a column of memory cells such as in a NAND string. Each circle represents a memory hole or a memory cell associated with the word line layer. Example columns of memory cells along a line **220** include C**0** to C**11**. Columns C**0**, C**3**, C**4**, C**7**, C**8** and C**11** represent the drain side columns of respective NAND strings. Columns C**1**, C**2**, C**5**, C**6**, C**9** and C**10** represent the source side columns of respective NAND strings. The figure represents a simplification, as many more rows of memory holes will typically be used, extending to the right and left in the figure. Also, the figures are not necessarily to scale. The columns of memory cells can be arranged in subsets such as sub-blocks.

Further, the NAND strings are arranged in sets, where each NAND string in a set has an SGD transistor with a common control gate voltage. See also FIG. **2**B. Regions **201**, **203**, **205**, **207**, **208** and **210** each represent a set of NAND strings, or a set of memory cells in a word line layer. For example, region **210** includes NAND strings NS**0**, . . . , NS**0**-**14**. A programming operation can involve one set of NAND strings. Each NAND string in a set can be associated with a respective bit line which is independently controlled to allow or inhibit programming.

The drawings are not to scale and do not show all memory columns. For example, a more realistic block might have twelve memory columns in the y direction as shown, but a very large number such as 32,000 memory columns in the x direction, for a total of 384,000 memory columns in a block. With U-shaped NAND strings, 192,000 NAND strings are provided in this example. With straight NAND strings, 384,000 NAND strings are provided in this example. Assuming there are twenty-four memory cells per column, there are 384,000×24=9,216,000 memory cells in the set.

FIG. **2**B depicts a top view of example select gate layer portions, consistent with FIG. **2**A. In one approach, the select gate layer **215** is different than a WLL in that a separate SGD layer portion or line, is provided for each set of NAND strings. That is, each single row of SGD transistors extending in the x direction is separately controlled. In other words, the control gates of the SGD transistors in each set of NAND strings are commonly controlled.

Further, an SGS layer portion or line is provided for a pair of rows of SGS transistors extending in the x direction, in one approach, for adjacent sets of NAND strings. Optionally, additional slits are used so that a separate SGS layer portion is provided for a single row of SGS transistors extending in the x direction. Thus, the control gates of the SGS transistors in a pair of rows of SGS transistors, or in a single row of SGS transistors, are also commonly controlled.

The SGS and SGD layer portions are created due to slits **239**, **240**, **241**, **242**, **243**, **245**, **247** and **248**. The slits extend partway down in the stack as depicted by example slit **241** in FIG. **2**C. Regions **227**, **228**, **229**, **232**, **233** and **237** represent SGD transistors in SGD layer portions **216**, **218**, **219**, **223**, **224** and **226**, respectively. Regions **253** and **254**, **255** and **257**, and **258** and **259** represent SGS transistors in SGS layer portions **217**, **221** and **225**, respectively. Regions **255** and **257**, **258** and **259**, represent SGS transistors in SGS layer portions **221** and **225**, respectively. The portion **209** from FIG. **2**A is repeated for reference.

The select gate transistors are associated with NAND strings NS**0**-NS**5**.

FIG. **2**C depicts an embodiment of a stack **231** showing a cross-sectional view of the portion **209** of FIG. **2**A, along line **220**. This example includes three select gate layers, SGL**1**, SGL**2** and SGL**3**. In this case, the slit extends down to DL**22**, so that three separate layers of select gate transistors are formed in each column of each NAND string. The stack has a top **287** and a bottom **238**.

The conductive layers of the select gates can have a same height (channel length) as the conductive layers of the memory cells, in one approach. This facilitates the fabrication of the memory device. In a column, the individual select gate transistors together are equivalent to one select gate transistor having a channel length which is the sum of the channel lengths of the individual select gate transistors. Further, in one approach, select gate transistors in a column (e.g., in layers SGL**1**, SGL**2** and SGL**3**) are connected and received a common voltage during operations. The SGS transistors can have a similar construction as the SGD transistors. Further, the SGS and SGD transistors can have a similar construction as the memory cell transistors.

The substrate may be p-type and can provide a ground which is connected to the top select gate layer, in one approach. A via **244** connects a drain side of C**0** and NS**0** to a bit line **288**. A via **262** connects a source side of C**1** and NS**0** to a source line **289**. Back gates **263**, **264**, **265** and **266** are provided in NS**0**, NS**1**, NS**2** and NS**3**, respectively.

Regions D**1**, D**2**, D**3** and D**4** represent SGD transistors and regions S**1**, S**2**, S**3** and S**4** represent SGS transistors, in SGL**1**.

FIG. **3**A depicts a top view of an example word line layer **304** of the block BLK**0** of FIG. **1**A, in a straight NAND string embodiment. In this configuration, a NAND string has only one column, and the source-side select gate is on the bottom of the column instead of on the top, as in a U-shaped NAND string. Moreover, a given level of a block has one WLL which is connected to each of the memory cells of the layer. Insulation-filled slits **346**, **347**, **348**, **349** and **350** can also be used in the fabrication process to provide structural support for the stack when undoped polysilicon layers are removed by a wet etch and a dielectric is deposited to form the alternating dielectric layers. A dashed line **305** extends through columns C**12**-C**17**. A cross-sectional view along line **305** of portion **307** is shown in FIG. **3**C.

Regions **340**, **341**, **342**, **343**, **344** and **345** represent the memory cells (as circles) of respective sets of NAND strings. For example, region **340** represents memory cells in NAND strings NS**0**A, . . . , NS**0**A-**14**. Additional NAND strings include NS**1**A, NS**2**A, NS**3**A, NS**4**A and NS**5**A.

Alternatively, the layer **304** represents an SGS layer, in which case each circle represents an SGS transistor.

FIG. **3**B depicts a top view of an example SGD layer **362**, consistent with FIG. **3**A. Slits **357**, **358**, **359**, **360** and **361** divide the SGD layer into portions **363**, **364**, **365**, **366**, **367** and **368**. Each portion connects the SGD transistors in a set of NAND strings. For example, SGD layer portion **363** or line connects the SGD transistors in the set of NAND strings NS**0**A to NS**0**A-**14**. Regions **351**, **352**, **353**, **354**, **355** and **356** represent the SGD transistors (as circles) of respective sets of NAND strings in the SGD layer portions **363**, **364**, **365**, **366**, **367** and **368**, respectively. The portion **307** from FIG. **3**A is also repeated. The select gate transistors are associated with NAND strings NS**0**A-NS**5**A.

FIG. **3**C depicts an embodiment of a stack **376** showing a cross-sectional view of the portion **307** of FIG. **3**A, along line **305**. In this example, three SGD layers, three SGS layers and dummy word line layers DWLL**1** and DWLL**2** are provided. Columns of memory cells corresponding to NAND strings NS**0**A-NS**3**A are depicted in the multi-layer stack. The stack includes a substrate **101**, an insulating film **250** on the substrate, and a portion of a source line SL**0**A. Additional straight NAND strings in a SGD line subset extend behind the NAND strings depicted in the cross-section, e.g., along the x-axis. NS**0**A has a source end SEa and a drain end DEa. The slits **346**, **347** and **348** from FIG. **3**A are also depicted. A portion of the bit line BLOA is also depicted. A conductive via **373** connects DEa to BL**0**A. The columns are formed in memory holes MH**0**-MH**4**. The memory holes are columnar and extend at least from a top **370** to a bottom **371** of the stack.

The source line SL**0**A is connected to the source ends of each NAND string. SL**0**A is also connected to other sets of memory strings which are behind these NAND strings in the x direction.

Word line layers, e.g., WLL**0**-WLL**23**, and dielectric layers, e.g., DL**0**-DL**24**, are arranged alternatingly in the stack. SGS transistors SGS**1***a*, SGS**1***b*, SGS**1***c* and SGS**1***d* are formed in the SGS**1** layer.

A region **246** of the stack is shown in greater detail in FIG. **4**A.

Regions SGD**1***a*, SGD**1***b*, SGD**1***c* and SGD**1***d* represent SGD transistors.

FIG. **4**A depicts a view of the region **246** of FIG. **3**C, showing memory cells MC**1** to MC**5**. A number of layers can be deposited along the sidewalls of the column and within each word line layer. These layers can include oxide-nitride-oxide (O—N—O) and polysilicon layers which are deposited, e.g., using atomic layer deposition. For example, the column includes a charge-trapping layer or film (CTL) **403** such as SiN or other nitride, a tunnel oxide (TNL) **404**, a polysilicon body or channel (CH) **405**, and a dielectric core (DC) **406**. A word line layer includes a block oxide (BOX) **402**, a block high-k material **401**, a barrier metal **400**, and a conductive metal such as W **399** as a control gate. For example, control gates CG**0**, CG**1**, CG**2**, CG**3** and CG**4** are provided for the memory cells MC**0**, MC**1**, MC**2**, MC**3** and MC**4**, respectively. In another approach, all of these layers except the metal are provided in the column. Additional memory cells are similarly formed throughout the columns. The layers in the memory hole form a columnar active area (AA) of the NAND string.

The use of one or more dummy memory cells between the select gate transistors and the data-storing memory cells is useful since program disturb can be greater for memory cells adjacent to, or close to, the select gate transistors. These edge cells have a lower amount of channel boosting due to constraints on the voltages of the select gate transistors of an inhibited NAND string.

When a memory cell is programmed, electrons are stored in a portion of the CTL which is associated with the memory cell. These electrons are drawn into the CTL from the channel, and through the TNL. The Vth of a memory cell is increased in proportion to the amount of stored charge. During an erase operation, the electrons return to the channel.

Each of the memory holes can be filled with a plurality of annular layers comprising a block oxide layer, a charge-trapping layer, a tunneling layer and a channel layer. A core region of each of the memory holes is filled with a body material, and the plurality of annular layers are between the core region and the WLLs in each of the memory holes.

As mentioned, parasitic cells can be formed between the memory cells during programming of the memory cells. For example, a parasitic cell PC**0_1** is formed between MC**0** and MC**1**, a parasitic cell PC**1_2** is formed between MC**1** and MC**2**, a parasitic cell PC**2_3** is formed between MC**2** and MC**3**, and a parasitic cell PC**3_4** is formed between MC**3** and MC**4**.

FIG. **4**B depicts a cross-section view of the region **246** of FIG. **4**A along line **444**. Each layer is ring-shaped in one possible approach, except the core filler, which is a cylinder.

FIG. **4**C depicts an expanded view of a portion **410** of the NAND string of FIG. **4**A, showing the formation of a parasitic cell PC**0_1** between the memory cells MC**0** and MC**1**. When a voltage is applied to the control gate of a memory cell via a respective word lines an electric field is generated. For MC**0**, the electric field includes a component **490** which causes electrons **471** to tunnel into a region **470** of the charge-trapping layer **403**, from the channel **405**. Additionally, fringing fields **492** and **491** are created. These are components of the electric field that are directed to a region of the charge-trapping layer which is not directly adjacent to the control gate. For example, the fringing field **491** causes an electrons **463** to tunnel into a region **460** of the charge-trapping layer **403**, from the channel **405**, thereby forming PC**0_1**. Similarly, for MC**1**, the electric field includes a component **480** which causes electrons **451** to tunnel into a region **450** of the charge-trapping layer **403**, from the channel **405**. A fringing field **482** causes electrons **461** and **462** to tunnel into the region **460**. An additional fringing field **481** causes a parasitic cell to be formed above MC**1**.

When a memory cell on a selected word line is subsequently read back, control gate read voltages such as VrA, VrB and VrC (FIG. **8**A) are applied to the memory cell while sensing circuitry determine whether the memory cell is in a conductive state. At the same time, a read pass voltage, Vr (e.g., 8-9 V), is applied to the remaining word lines. Since the control gate read voltage is typically significantly less than the read pass voltage, the parasitic cells adjacent to the selected word line will not be fully conductive during the read operation. This results in the memory cell appearing to have an upshifted Vth. The other parasitic cells, which are not adjacent to the selected word line, will be conductive since they are between word lines which both receive Vr.

FIG. **4**D depicts a table showing a magnitude of an electric field (Efield) which is seen by the parasitic cell PC**0_1** of FIG. **4**C with various combinations of pass voltages and data states on MC**0** and MC**1**. Assume that the word line programming order proceeds from the source-side word line, WLL**0**, to the drain side word line WLL**22**, one word line at a time. In this case, MC**0** is programmed before MC**1**. MC**0** is on a source side of MC**1**, and WLL**0** is on a source side of WLL**1**. MC**1** is on a drain side of MC**0**, and WLL**1** is on a drain side of WLL**0**. The example uses verify voltages of VvA=1, 2 and 3 V for the A, B and C states, respectively. Also, a final Vpgm of 21, 22 and 23 V is used for the A, B and C states, respectively, on average. For a given data state, some cells will program faster and have a lower final Vpgm and some cells will program slower and have a higher final Vpgm. PC**0_1** is subject to an electric field a first time based on the Vpgm on MC**1** and the Vpass on MC**1**, and a second time based on the Vpgm on MC**1** and the Vpass on MC**0**.

During a programming operation, a sequence occurs in which the A, B and C state memory cells pass their respective verify test at a first, second or third phase of the programming operation. See also FIG. **10**A. As a result, the pass voltage on the later-programmed word line can be set as a function of the phase of the programming operation. In particular, the pass voltage can be set at a maximum level, intermediate level or minimum level in the first, second or third phases, respectively. This approach can be modified such as to include multiple intermediate levels.

The table identifies the various combinations which are possible in an asymmetric word line bias technique. In the first three rows of the table, when PC**0**_**1** is subject to an electric field a first time and the final Vpgm for MC**0** is 21 V, Vpass on MC**1** is set to 9 V so that the total electric field (Efield) is 30 V. Subsequently, three options are possible when PC**0**_**1** is subject to an electric field a second time. If MC**1** is programmed to the A, B or C state, the total Efield will be 28, 29 or 30 V, respectively. In each case, a lower pass voltage of 7 V is used for MC**0** so that the total Efield is less than or equal to, but not higher than, the total Efield when PC**0**_**1** is subject to an electric field the first time. As a result, due to the use of a lower Vpass on the previously-programmed cell (MC**0**) when programming a currently selected cell (MC**1**), the parasitic cell (PC**0**_**1**) between the previously-programmed cell and the currently selected cell is not further programmed. This advantageously means the Vth of MC**0** does not change when MC**0** is read back.

Each memory cell experiences this benefit in turn as it is programmed in the word line order.

In the next three rows of the table, when PC**0**_**1** is subject to an electric field a first time and the final Vpgm for MC**0** is 22 V, Vpass on MC**1** is set to 8 V so that the total electric field (Efield) is again 30 V. If MC**1** is programmed to the A, B or C state, the total Efield will be 28, 29 or 30 V, respectively. In each case, a lower pass voltage of 7 V is used for MC**0** so that the total Efield is less than or equal to, but not higher than, the Efield when PC**0**_**1** is subject to an electric field the first time.

In the last three rows of the table, when PC**0**_**1** is subject to an electric field a first time and the final Vpgm for MC**0** is 23 V, Vpass on MC**1** is set to 7 V so that the total electric field (Efield) is again 30 V. If MC**1** is programmed to the A, B or C state, the total Efield will be 28, 29 or 30 V, respectively. In each case, a lower pass voltage of 7 V is used for MC**0** so that the total Efield is less than or equal to, but not higher than, the Efield when PC**0**_**1** is subject to an electric field the first time.

Note that when programming occurs for a word line which is not adjacent to PC**0**_**1**, PC**0**_**1** will be subject to an electric field which is based on pass voltages on the adjacent word lines. This electric field will be lower than any electric field it experiences when PC**0**_**1** is adjacent to a currently programmed word line since a pass voltage is lower than any final program voltage.

As the memory device accumulates program-erase cycles, the programming speed of the memory cells may increases so that the final Vpgm becomes lower. However, this is expected to occur uniformly for the different data states so that the same pass voltages as used in a fresh memory device will continue to produce the same result of avoiding or minimizing the programming of parasitic cells.

FIG. **5**A depicts a cross-sectional view in a word line direction of memory cells comprising a flat control gate and charge-trapping regions a 2D example of memory cells in the memory structure **126** of FIG. **1**B. This is one alternative to a 3D memory device in which the pass voltages can be optimized during programming to reduce program disturb. Charge-trapping memory can be used in NOR and NAND flash memory device. This technology uses an insulator such as an SiN film to store electrons, in contrast to a floating-gate MOSFET technology which uses a conductor such as doped polycrystalline silicon to store electrons. As an example, a word line (WL) **524** extends across NAND strings which include respective channel regions **506**, **516** and **526**. Portions of the word line provide control gates **502**, **512** and **522**. Below the word line is an inter-poly dielectric (IPD) layer **528**, charge-trapping layers **504**, **514** and **521**, polysilicon layers **505**, **515** and **525** and tunnel oxide (TOx) layers **509**, **507** and **508**. Each charge-trapping layer extends continuously in a respective NAND string.

A memory cell **500** includes the control gate **502**, the charge-trapping layer **504**, the polysilicon layer **505** and a portion of the channel region **506**. A memory cell **510** includes the control gate **512**, the charge-trapping layer **514**, a polysilicon layer **515** and a portion of the channel region **516**. A memory cell **520** includes the control gate **522**, the charge-trapping layer **521**, the polysilicon layer **525** and a portion of the channel region **526**.

Further, a flat control gate may be used instead of a control gate that wraps around a floating gate. One advantage is that the charge-trapping layer can be made thinner than a floating gate. Additionally, the memory cells can be placed closer together.

FIG. **5**B depicts a cross sectional view along line **559** in FIG. **5**A, showing a NAND string **530** having a flat control gate and a charge-trapping layer. The NAND string **530** includes an SGS transistor **531**, example storage elements **500**, **532**, . . . , **533** and **534**, and an SGD transistor **535**. The memory cell **500** includes the control gate **502** and an IPD portion **528** above the charge-trapping layer **504**, the polysilicon layer **505** and the tunnel oxide layer **509**. The memory cell **532** includes a control gate **536** and an IPD portion **537** above the charge-trapping layer **504**, the polysilicon layer **505** and the tunnel oxide layer **509**.

The control gate layer may be polysilicon and the tunnel oxide layer may be silicon oxide, for instance. The IPD layer can be a stack of high-k dielectrics such as AlOx or HfOx which help increase the coupling ratio between the control gate layer and the charge-trapping or charge storing layer. The charge-trap layer can be a mix of silicon nitride and oxide, for instance. A difference between a floating gate memory cell and the flat memory cell is the height of the charge storage layer. A typically floating gate height may be about 100 nm, while a charge-trap layer can be as small as 3 nm, and the polysilicon layer can be about 5 nm.

The SGD and SGS transistors have the same configuration as the storage elements but with a longer channel length to ensure that current is cutoff in an inhibited NAND string.

FIG. **5**C depicts an expanded view of a portion **540** of the NAND string of FIG. **5**B, showing the formation of a parasitic cell PC**1**. The charge-trapping layer **504** includes regions **541** and **543** which are directly under and adjacent to the memory cells **500** and **532**, respectively, and a region PC**1** which forms a parasitic cell due to a fringing electric field **544** and **545** from the memory cells **500** and **532**, respectively.

FIG. **6**A depicts a plot showing an increase in the Vth of a memory cell on a selected word line WLn due to program disturb, as a function of a data state of a memory cell on a later-programmed word line, WLn+1. The solid lines represent symmetric boosting in which an equal pass voltage is used on the word lines which are adjacent to a word line which is being programmed throughout the programming operation. The dashed lines represent asymmetric boosting in which modified pass voltages are used on the word lines which are adjacent to a word line which is being programmed, such as described in connection with FIG. **4**D. The modified pass voltages are set based on the observation that a parasitic cell is initially programmed based on a first combination of a pass voltage and a final Vpgm, and that further programming can be avoided by ensuring that a second combination of pass voltage and a final Vpgm does not provide a higher electric field than the first combination.

For the case of symmetric boosting, after the memory cell on WLn is programmed to a data state, the memory cell experiences an increase in its Vth after the memory cell on WLn+1 is programmed to a higher data state. This applies for memory devices with large and small spaces between word lines and regardless of the existence of parasitic cells.

Referring to plot **603**, with symmetric boosting, a memory cell on WLn which remains in the erased (E) state experiences an increase in Vth when the memory cell on WLn+1 is programmed to the A, B or C state (but not when the memory cell on WLn+1 remains in the Er state).

Referring to plot **602** and **602***a*, with symmetric boosting, a memory cell on WLn which is programmed to the A state experiences an increase in Vth when the memory cell on WLn+1 is programmed to the B or C state (but not when the memory cell on WLn+1 is programmed to the A state or remains in the erased (E) state).

Referring to plot **601** and **601***a*, with symmetric boosting, a memory cell on WLn which is programmed to the B state experiences an increase in Vth when the memory cell on WLn+1 is programmed to the C state (but not when the memory cell on WLn+1 is programmed to the A state or remains in the erased (E) state).

The slope of the increase in the Vth of the memory cell on WLn is higher when the spacing between word lines is smaller.

Referring to plot **600**, with symmetric or asymmetric boosting, a memory cell on WLn which is programmed to the C state does not experience an increase in Vth when the memory cell on WLn+1 is programmed, since the memory cell on WLn+1 cannot be programmed to a higher state, since the C state is the highest state in this example.

For the case of asymmetric boosting, after the memory cell on WLn is programmed to a data state, the memory cell does not experience an increase in its Vth even after the memory cell on WLn+1 is programmed to a higher data state. This is depicted by the plots **602***b* and **601***b* for the A and B states, respectively, of the memory cell on WLn. Referring to plot **603***a*, a memory cell in the Er state on WLn will experience a reduced amount of upshift in Vth with asymmetric boosting compared to symmetric boosting. As mentioned, a cell in the Er state is not programmed so that it cannot compensate for the increase in Vth in adjacent parasitic cells on either side of the Er state cell. However, the increase in the Vth of the parasitic cell between WLn and WLn+1 can be decreased by the use of asymmetric boosting during programming as described herein. A further measure to account for the increase in Vth for Er state cells is provided by a read operation as discussed herein.

FIG. **6**B depicts a plot showing an increase in the Vth of an erased state memory cell on a selected word line WLn as a function of a data state of a memory cell on a previously-programmed word line, WLn−1. The plot **604** with a solid line shows the case of symmetric boosting and the plot **604***a* with the dashed line shows the case of asymmetric boosting. A parasitic cell affects the Vth of an erased state cell on WLn similarly regardless of whether the parasitic cell is between WLn and WLn+1, or between WLn and WLn−1. The total upshift of the erased state cell is equal to the sum of plots **603** and **604** with symmetric boosting, or the sum of plots **603***a* and **604***a* with asymmetric boosting.

FIG. **7** depicts a circuit diagram of a NAND string consistent with the memory devices of FIGS. **2**C, **3**C and **5**B. An example NAND string NS**0**A includes SGD transistors **701**, **702** and **703**, a drain-side dummy memory cell **704**, data-storing memory cells **705**, . . . , **706**, **707**, **708**, . . . , **709**, a source-side dummy memory cell **710**, and SGS transistors **711**, **712** and **713**. A bit line **715** connects the drain end of the NAND string to sensing circuitry **700**, which is used to sense the NAND string during operations involving the select gate transistors and the memory cells. A source line **714** is connected to a source end of the NAND string. Voltage drivers can be used to provide the voltages depicted. For example, Vsg is applied to the control gates of the SGD transistors, which are connected to one another and to the control gates of the SGS transistors, which are connected to one another. Vsg can also be applied to the dummy memory cells **704** and **710**. Vbl is the bit line voltage and Vsl is the source line voltage.

A similar circuit diagram can be provided for the NAND string **530** of FIG. **5**B by omitting the extra select gate transistors and the dummy memory cells.

In this example, the memory cell **707** is selected for programming and is connected to a selected word line WLLn which receives Vpgm. The memory cell **708** is an adjacent, previously-programmed memory cell which is connected to a word line WLLn−1 which receives Vpass**2**. The memory cell **706** is an adjacent, later-programmed memory cell which is connected to a word line WLLn+1 which receives Vpass**1**. The remaining word lines, which are not adjacent to WLLn, including WLL**0** to WLLn−2 and WLLn+2 to WLL**22**, can receive a nominal pass voltage, Vpass_nom, which can be optimized separately from Vpass**1** and Vpass**2**.

FIG. **8**A depicts threshold voltage (Vth) distributions of memory cells in an example embodiment in which there is an erased state and three target data states (e.g., programmed states). The horizontal axis depicts Vth and the vertical axis depicts a number of memory cells. The memory cells initially have an erased state Vth distribution **800** which is associated with a verify voltage of Vv_Er, and are subsequently programmed to the A, B or C target data states as represented by Vth distributions **810**, **820** and **830**, respectively, using verify voltages of VvA, VvB and VvC, respectively. Read levels of VrA, VrB and VrC are also depicted. A pass voltage Vpass can be applied to all unselected word line during a read operation, including a verify operation during a programming operation as well as a read back of data after a programming operation is completed. This figure represents an ideal situation without program disturb.

FIG. **8**B depicts widening of the Vth distributions of FIG. **8**A due to program disturb, in a symmetric boosting case. In this case, the pass voltage on WLn+1 (Vpass**1**) equals the pass voltage on WLn−1 (Vpass**2**) during a programming operation. This is referred to as symmetric boosting. Program disturb typically increases the upper tail of a Vth distribution. The Vth distributions **800**, **810** and **820** transition to the distributions **801**, **811** and **821**, respectively. The C state generally does not experience program disturb.

FIG. **8**C depicts a reduction in the widening of the Vth distributions of FIG. **8**A due to program disturb, in an asymmetric boosting case. In this case, the pass voltage on WLn+1 (Vpass**1**) initially exceeds the pass voltage on WLn−1 (Vpass**2**) and subsequently decreases during a programming operation. See also FIG. **10**B. The Vth distribution **804** is wider than the Vth distribution **800** but by a smaller amount than an amount by which the Vth distribution **801** is wider than the Vth distribution **800** in FIG. **8**B. The Vth distributions for the target data states are advantageously not widened.

However, the Vth distribution is still widened for the Er state. A Vth distribution **804** represents the total amount of widening for the Er state and is based on a Vth distribution **802** which is due to the WLn+1 memory cells and a Vth distribution **803** which is due to the WLn−1 memory cells. Typically, the program disturb which is caused by the WLn−1 memory cells is greater than the program disturb which is caused by the WLn+1 memory cells.

FIG. **8**D depicts a plot showing a reduction in the Vth upshift for different data states on WLn as a function of dVpass=Vpass**1**−Vpass**2**. The horizontal axis depicts dVpass and the vertical axis depicts an upshift or widening of a Vth distribution of a set of memory cells on WLn after programming of a set of memory cells on WLn+1. Plots **840-843** indicate that the Vth upshift for an Er state cell decreases as Vpass increases and levels off at a non-zero value. Furthermore, the Vth upshift is a function of the data state on WLn−1. The plots **840**, **841**, **842** and **843** represent the Vth upshift for an Er state memory cell on WLn when the corresponding memory cell on WLn−1 on the same charge-trapping layer is in the C, B, A or Er state, respectively. Plots **844** and **845** indicate that the Vth upshift for an A or B state cell, respectively, decreases to zero as dVpass increases to an optimum value of dVpass_max, consistent with FIG. **8**C. Plot **846** indicates that the Vth upshift for a C state cell is essentially zero. dVpass_max may be about 2-4 V, for example.

FIG. **8**E depicts an example in which different values of a read voltage are used to account for different levels of disturb to memory cells in the erased state on WLn as a function of four possible data states of memory cells on WLn−1. The ideal, undisturbed Vth distributions **800** and **810** for the Er and A states, respectively, of FIG. **8**A, are repeated. The Vth distributions **800**, **850**, **851** and **852** represent the Er state memory cells on WLn which are adjacent to an Er, A, B or C state memory cell, respectively, on WLn−1. In this case, the Vth upshift is approximately linear with the increase in the data state on WLn−1. Alternatively, the Vth distributions **800**, **850**, **851** and **852** represent the Er state memory cells on WLn which are adjacent to memory cell in a first, second, third or fourth Vth range, respectively, on WLn−1. In this case, the Vth upshift is approximately linear with the increase in the Vth range on WLn−1.

A read voltage is depicted for each data state or Vth range on WLn−1. For example, VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3** and VrA_C/R**4**, are optimized read voltages for distinguishing between the Er and A states for a memory cell on WLn when the memory cell on the same charge-trapping layer on WLn−1 is in the Er, A, B or C, state respectively, or in a Vth range R**1**, R**2**, R**3** or R**4**, respectively. See FIGS. **13**A and **13**B for related examples of Vth ranges.

In one approach, the read voltages are optimized by testing of various memory devices and stored in the storage region **113** of FIG. **1**B. A read voltage can be at a location which is midway between the upper end of the Vth distribution of the Er state and the lower end of the Vth distribution of the A state, if these Vth distributions do not overlap. For example, VrA_Er/R**1** is midway between the upper end of the Vth distribution of the Er state (at Vv_Er) and the lower end of the Vth distribution of the A state (at VvA).

In another approach, a read voltage can be at a location which is a minimum between overlapping Vth distributions of the Er state and the A state. For example, VrA_A/R**2** is at a minimum (e.g., in a valley) between the Vth distribution **850** of the Er state and the Vth distribution **810** of the A state. Also, VrA_B/R**3** is at a minimum between the Vth distribution **851** of the Er state and the Vth distribution **810** of the A state. Also, VrA_C/R**4** is at a minimum between the Vth distribution **852** of the Er state and the Vth distribution **810** of the A state.

FIG. **8**F depicts another example in which different values of a read voltage are used to account for different levels of disturb to memory cells in the erased state on WLn as a function of four possible data states of memory cells on WLn−1. The ideal, undisturbed Vth distributions **800** and **810** for the Er and A states, respectively, of FIG. **8**A, are repeated. The Vth distributions **800**, **860**, **861** and **862** represent the Er state memory cells on WLn which are adjacent to an Er, A, B or C state memory cell, respectively, on WLn−1. In this case, the Vth upshift is nonlinear with the increase in the data state on WLn−1. In particular, the increase in Vth is relatively larger when the data state on WLn−1 is relatively higher. Alternatively, the Vth distributions **800**, **860**, **861** and **862** represent the Er state memory cells on WLn which are adjacent to memory cell in a first, second, third or fourth Vth range, respectively, on WLn−1. In this case, the increase in Vth is relatively larger when the Vth range on WLn−1 is relatively higher.

A read voltage is depicted for each data state or Vth range on WLn−1. For example, VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3** and VrA_C/R**4**, are optimized read voltages for distinguishing between the Er and A states for a memory cell on WLn−1 when the memory cell on the same charge-trapping layer on WLn−1 is in the Er, A, B or C, state respectively, or in a Vth range R**1**, R**2**, R**3** or R**4**, respectively. See FIGS. **13**A and **13**B for related examples of Vth ranges.

FIG. **8**G depicts Vth distributions of memory cells in an example embodiment in which there is an erased state and seven target data states. The horizontal axis depicts Vth and the vertical axis depicts a number of memory cells. The memory cells initially have an erased state Vth distribution **870** which is associated with a verify voltage of Vv_Er, and are subsequently programmed to the A, B, C, D, E, F and G target data state as represented by Vth distributions **871**, **872**, **873**, **874**, **875**, **876** and **877**, respectively, using verify voltages of VvA, VvB, VvC, VvD, VvE, VvF and VvG, respectively. This figure represents an ideal situation without program disturb. Due to symmetric boosting, the Vth distributions **800**, **871**, **872**, **873**, **874**, **875** and **876** transition to the distributions **880**, **881**, **882**, **883**, **884**, **885** and **886**, respectively. The C state generally does not experience program disturb.

FIG. **8**H depicts an example in which different values of a read voltage are used to account for different levels of disturb to memory cells in the erased state on WLn as a function of eight possible data states of memory cells on WLn−1, consistent with FIG. **8**G. The ideal, undisturbed Vth distributions **870** and **871** for the Er and A states, respectively, of FIG. **8**G, are repeated. The Vth distribution **800** represents the Er state memory cells on WLn which are adjacent to an Er, A or B state memory cell on WLn−1. The Vth distribution **890** represents the Er state memory cells on WLn which are adjacent to a C or D state memory cell on WLn−1. The Vth distributions **891**, **892** and **893** represent the Er state memory cells on WLn which are adjacent to E, F and G state memory cells, respectively, on WLn−1.

A read voltage is depicted for each data state or Vth range on WLn−1. For example, VrA_Er,A,B is an optimized read voltage for distinguishing between the Er and A states for a memory cell on WLn−1 when the memory cell on the same charge-trapping layer on WLn−1 is in the Er, A or B state. VrA_C,D is an optimized read voltage for distinguishing between the Er and A states for a memory cell on WLn−1 when the memory cell on the same charge-trapping layer on WLn−1 is in the C or D state. VrA_E, VrA_F and VrA_G are optimized read voltages for distinguishing between the Er and A states for a memory cell on WLn−1 when the memory cell on the same charge-trapping layer on WLn−1 is in the E, F or G state, respectively.

This approach assumes that groups of memory cells on WLn−1 with adjacent data states cause a similar disturb on an Er state memory cell on WLn. Read time is reduced by reducing the number of read voltages. Other groupings of data states could be used.

FIG. **9**A depicts a process for programming and then reading memory cells in a memory device. The memory cells are programmed according to a word line programming order using asymmetric boosting (step **900**). The memory cells of a first word line are programmed, then the memory cells of a second word line are programmed, and so forth. In one approach for a 3D memory device, word line layers are programmed starting from a bottom layer and continuing layer-by-layer toward the top layer. In some case, programming extends from the source side to the drain side of a set of NAND strings. In some cases, not all layers are used so that some layers remain in the erased state. For example, the programming can provide the selected memory cells in four, eight or sixteen data states. Step **901** involves reading the memory cells on a selected word line (WLn) with compensation for the erased state memory cells due to program disturb from memory cells on WLn−1. WLn+1 is a drain-side word line and WLn−1 is a source side word line of WLn.

The read operation can be implemented, e.g., at a system level or a memory chip level. In a system level approach, a controller such as a state machine which is external to the memory chip is used to issue commands such as each control gate read level and to toggle out data such as the result of sensing operations. Firmware of the controller can be configured to provide the desired functionality without modifying the memory chip. In the chip level approach, much of the functionality is provided using circuits on the memory chip. This approach can provide faster reads but consumes space on the memory chip.

The read operation thus involves looking back at the data states of the previously-programmed memory cells on WLn−1, after programming using asymmetric boosting.

FIG. **9**B depicts an example programming operation for memory cells of a selected word line WLn which minimizes program disturb by optimizing pass voltages, consistent with step **910** of FIG. **9**A. Step **910** begins a programming operation for a set of memory cells connected to a set of word lines. Step **911** selects a word line (WLn) for programming. Step **912** initializes Vpgm. Step **913** sets Vpass**1** on WLn+1 and Vpass**2** on WLn−1. This step can be based on various criterion. For example, step **913***a* determines a current program loop number. Vpass**1** can be relatively higher when the current program loop number is relatively lower. Vpass**1** can be initially fixed at a maximum level and subsequently made progressively smaller during a programming operation. Vpass**1** can begin to become progressively smaller when the current program loop number exceeds a threshold count. See FIGS. **10**A and **10**B. This approach can be implemented by storing a table of program loop number cross-referenced to Vpass**1** and, optionally, Vpass**2**, so that a pass voltage for each program loop is predetermined.

Step **913***b* determines a current data state being verified. Vpass**1** can be relatively higher when the current data state being verified is relatively lower. This approach can be implemented by storing a table of data state cross-referenced to Vpass**1** and, optionally, Vpass**2**. This approach is adaptive since the specified program loops which are used to verify a given data state can vary based on factors such as programming speed. For example, Vpass**1** can be initially fixed at a maximum level when the A state cells are being verified and subsequently made progressively smaller when the B state cells are being verified.

Step **913***c* comprises monitoring a programming progress of the memory cells on WLn. For example, a reduction in Vpass**1** can be triggered when on all, or a specified majority, of the A state cells pass their verify test. A further reduction in Vpass**1** can be triggered based on all, or a specified majority, of the B state cells passing their verify test. See FIG. **10**B. Or, further multiple reductions in Vpass**1** can be triggered based on different portions (e.g., 25%, 50%, 75%) of the B and/or C state cells passing their verify tests. This approach can be implemented by storing a table of threshold counts of memory cells which pass a verify test cross-referenced to Vpass**1** and, optionally, Vpass**2**. In each program-verify iteration, a count can be obtained of a number of cells which pass the verify test for one or more data states which are subject to a verify test. A decision to adjust Vpass**1** (or Vpass**2**) can be made based on whether the count exceeds the threshold counts.

Step **914** concurrently applies Vpgm to the selected word line (WLn), Vpass**1** to the adjacent later-programmed word line (WLn+1), Vpass**2** to the adjacent previously-programmed word line (WLn−1) and Vpass_nom to the remaining word lines. Step **915** determines one or more current data states to verify and performs the corresponding verify test or tests. For example, in FIG. **10**A, the A state cells are verified at program loops **1-4**, the B state cells are verified at program loops **4-7** and the C state cells are verified at program loops **7-10**.

Decision step **916** determines if the verify test or tests are passed for the current data state or states. For example, a verify test for a data state may be passed when all, or a significant majority, e.g., 90-95%, of the memory cells which are to be programmed to that data state have passed the verify test. A memory cell passes a verify test when a verify voltage is applied to its control gate via a word line and the memory cell is determined by sensing circuitry to be in a non-conductive state. In this case, the Vth of the memory cell exceeds the verify voltage. If decision step **916** is false, a next program loop (program-verify iteration) is performed. In this case, Vpgm is incremented at step **920** and step **913** is repeated. If decision step **916** is true, decision step **917** determines whether programming has been completed for the word line. If decision step **917** is false, a next program loop is performed and Vpgm is incremented at step **910**. If decision step **917** is true, a decision step **918** determines whether there is another word line to program. If decision step **918** is false, the programming operation ends at step **919**. If decision step **918** is true, a new word line is selected to be programmed at step **911**.

FIG. **10**A depicts a plurality of program-verify iterations consistent with the programming operation of FIG. **9**B. The horizontal axis depicts program loop number, and the vertical axis depicts control gate or word line voltage on WLn, the selected word line. Generally, a programming operation includes successive program-verify iterations. Each program-verify iteration has a program portion in which a program voltage is applied to the control gates of the memory cells via a selected word line, followed by a verify portion in which one or more verify voltages are applied to the control gates of the memory cells while at least some of the memory cells are sensed.

The voltage waveform **1000** depicts a series of program voltages or program pulses PP**1** to PP**10** and verify voltages. One or two verify voltages are provided after each program voltage, as an example, based on the programming phase and the corresponding target data states which are being verified. 0 V may be applied to the selected word line between the program and verify voltages. For example, during a first set of program loops **1010** (program loops **1-4**), e.g., one or more initial program-verify iterations, the A state cells are subject to a verify test using VvA as indicated by an example verify pulse VPa. During a second set of program loops **1020** (program loops **4-7**), e.g., one or more intermediate program-verify iterations, the B state cells are subject to a verify test using VvB as indicated by an example verify pulse VPb. During a third set of program loops **1030** (program loops **7-10**), e.g., one or more final program-verify iterations, the C state cells are subject to a verify test using VvC as indicated by an example verify pulse VPc. Loop **10** represents the final program-verify iteration. The program voltages are stepped up in each program loop after the first loop. This is an example of incremental step pulse programming (ISPP).

Further, the verifying of the A and B states overlaps in program loop **4**, where the A and B state cells are subject to a verify test using VvA and VvB, respectively, as indicated by an example waveform VPab. The verifying of the B and C states overlaps in program loop **7**, where the B and C state cells are subject to a verify test using VvB and VvC, respectively, as indicated by an example waveform VPbc.

The one or more intermediate program-verify iterations and the one or more final program-verify iterations are part of subsequent program-verify iterations which are after the one or more initial program-verify iterations.

In one approach, the data states are verified in predetermined program loops (e.g., the A state, A and B states, B state, B and C states and C state are verified in program loops **1-3**, **4**, **5-6**, **7** and **8-10**, respectively). In another approach, the verification of cells of a data state begins at a predetermined program loop but the verification continues until a program loop in which all, or nearly all, of the memory cell cells of that data state have passed the respective verify test (e.g., verification for the A state, B state and C state begins in program loops **1**, **4** and **7**, respectively). In another approach, the verification of cells of a data state begins based on the programming progress of cells of a lower data state. For example, the verification of cells of the B state may begin when a specified portion (e.g., 80%) of the memory cells of the A state have passed their respective verify test (e.g., 80% of the A state cells pass a verify test in program loop **3** so that verification of the B state begins in program loop **4**). Other approaches are possible as well.

Generally, a majority of the memory cells of a given data state will pass their verify test in a small range of program loops, while the slower cells of the given data state will pass their verify test at a later program loop and the faster cells of the given data state will pass their verify test at an earlier program loop. For example, a majority of the A state cells will pass their verify test at program loops **2** and **3**, while the slower A state cells will pass their verify test at program loop **4** and the faster A state cells will pass their verify test at program loop **1**.

The above techniques can be extended to cases where there are additional data states, e.g., eight or sixteen data states. For example, with eight data states A-G, several phases are possible in a programming operation. For instance, five phases may be used which include these groups of states which are verified concurrently: ABC, BCD, CDE, DEF and EFG. Another example is: ABCD, CDEF and EFG. Many other examples are possible.

FIG. **10**B depicts one example of pass voltages consistent with FIG. **10**A in which a decreasing level of Vpass**1** is used on WLn+1 and a fixed level of Vpass**2** is used on WLn−1. The horizontal axis represents a loop number and are aligned with the horizontal axis of FIG. **10**A. The vertical axis depicts a pass voltage level, and shows the verify voltages VvA, VvB and VvC and the nominal pass voltage of Vpass_nom. Vpass_nom can be equal to Vpass**1**_max, in one approach but, generally can be optimized independently of Vpass**1** and Vpass2.

Vpass**1** is initially at a maximum level, Vpass**1**_max (line **1040**), during the initial program-verify iterations, while the A state cells are being verified. Vpass**1** then steps down to an intermediate level, Vpass**1**_int (line **1041**), during the one or more intermediate program-verify iterations, while the B state cells are being verified. Vpass**1** can step down at this time because the programming of all or most of the A state cells has been completed such that the associated NAND string, including the parasitic cells, are inhibited from further programming while remaining B and C state cells on the selected word line continue to be programmed. The difference between Vpass**1**_int and Vpass**2** is sufficient to avoid program disturb of B state cells.

Vpass**1** then steps down to a final, minimum level, Vpass**1**_min (line **1042**), during the one or more final program-verify iterations, while the C state cells are being verified. Vpass**1** can step down to a minimum at this time because the programming of all or most of the B state cells has been completed. Vpass**1**_min can be the same as, or slightly higher than, Vpass**2**, for instance, since the C state cells are generally resistant to Vth widening. Vpass can be kept at the fixed minimum during a plurality of final program-verify iterations of the plurality of program-verify iterations (program loops **1-10**). Vpass**1** can be kept at the fixed minimum during at least two of the program voltages.

Vpass**2** may be held at a fixed level (line **1043**) which is equal to, or close to, Vpass_min. This approach is desirable because Vpass**1** is provided at the minimum level while Vpgm reaches its highest levels. The total electric field, which is proportional to Vpass**1**+Vpgm, is therefore smaller than if Vpass**1** was kept at Vpass**1**_max. As a result, the programming of the parasitic cells adjacent to the Er state cells especially is reduced.

In this example, the subsequent program-verify iterations (program loops **5-10**) comprise program-verify iterations (loops **5-7**) in which the first pass voltage is fixed at a first reduced level (Vpass**1**_int) followed by program-verify iterations (loops **8-10**) in which the first pass voltage is fixed at a second reduced level (Vpass**1**_min), below the first reduced level.

In one approach, the first pass voltage and the difference (dVpass) start to become progressively smaller when a determination is made that a specified portion of memory cells which are to be programmed to a lowest target data state (e.g., 90-95% of the A state cells) of the plurality of target data states have passed a verify test. For example, 95% or more of the A state cells may pass their verify test in program loop **4** so that Vpass**1** becomes smaller in program loop **5**.

In another approach, the first pass voltage and the difference start to become progressively smaller after a predetermined number of program-verify iterations (e.g., four) of the plurality of program-verify iterations have been performed.

Generally, the initial value of Vpass**1**, e.g., Vpass1_max, should exceed the final value of Vpass**2** by a specified amount. This applies when Vpass**2** is fixed, increasing or decreasing. This specified amount should be the difference (Diff**1**) in the verify levels of the highest and lowest programmed data states, e.g., VvA-VvC. This difference is known since the verify levels are configured in the control of the memory device. Typically, Diff**1** is the same as the difference (Diff**2**) between the final Vpgm levels of the highest and lowest programmed data states. Diff**2** can be determined by testing of the memory device. However, it is possible that

Diff1 and Diff2 are different, in which case the initial value of Vpass1 should exceed the final value of Vpass2 by and Diff2 since the parasitic cells are a function of the program voltages.

FIG. 11 depicts an example of a reading operation consistent with step 901 of FIG. 9A. Step 1100 involves receiving a read command to read selected memory cells of WLn. For example, the command may be issued by the host, outside the memory device, to the controller 122 and from the controller to the state machine. In other cases, the command is issued from within the memory device, in the controller. The state machine can then communicate a read command to the sense blocks, identifying the word line to be read and optionally, a specific group or page of memory cells within the word line.

Step 1101 involves reading unselected memory cells of WLn−1 to determine a data state (Vth range) for each unselected memory cell, and stores the read results such as in latches. For example, latches can be used which have the capability of storing data which is read from WLn−1 and WLn. Specifically, for a set of memory cells to be read on WLn, a corresponding set of adjacent memory cells are read on WLn−1. The set of memory cells to be read on WLn can be all, or only a portion, of all memory cells on WLn. Each Vth range can encompass part of a data state, exactly one data state, or multiple data states, or a combination of one or more data states and a part of a data state. For example, see FIGS. 13A, 13B, 14A and 14B.

Step 1102 involves reading the selected memory cells based on the read of WLn−1 to distinguish between the erased state (Er) and the lowest target data state (A). This read operation can comprise multiple read processes, where each read process can optimally distinguish between the Er and A states on WLn based on one of the data states or Vth ranges on WLn−1.

Step 1103 involves reading the selected memory cells based on the read of WLn−1 to distinguish between pairs of the adjacent target data states. For example, one read process distinguishes between the A and B states and another read process distinguishes between the B and C states, when there are three target data states. At step 1104, the read operation has completed.

FIG. 12A depicts a reading process consistent with step 1102 of FIG. 11. Step 1200 selects a data state (or Vth range) of the data states (or Vth ranges) for the memory cells on WLn−1. For example, this can be the Er, A, B or C state in a four-state memory, or one Vth range of multiple Vth ranges.

Step 1201 applies a control gate voltage which is a function of the data state (or Vth range) to WLn while sensing whether the selected memory cells are in a conductive state. For example, referring to see FIG. 8E, the selected data state at step 1200 can be the Er state in which case the control gate voltage at 1201 is VrA_Er/R1.

At step 1202, for each select memory cell on WLn, we select the sensing result if the adjacent unselected memory cell on WLn−1 has the data state (or Vth range) selected at step 1200. Thus, we select a result (SR1*a*-SR1*d*) from one read process of a set of read processes based on the threshold voltage range of another (WLn−1) memory cell. This sensing result is optimal for the selected memory cell. For example, if the Er state is selected at step 1200, the sensing result is optimal when the adjacent memory cell is in the Er state and is not optimal when the adjacent memory cell is in the A, B or C state. The non-optimal sensing results are discarded for the particular memory cell.

Decision step 1203 determines if there is a next data state (or Vth range) on WLn−1 to consider. If this decision step is true, step 1200 is repeated using the next control gate voltage (e.g., VrA_A/R2, VrA_B/R3 or VrA_C/R4). If decision step 1203 is false, the process ends at step 1204.

The data which is read from WLn can be transferred from latches and out to the state machine, controller and host.

WLn can be considered to have one memory cell, e.g., MC1, of a set of NAND strings, e.g., NS0A-NS3A, which is connected to one word line, e.g. WL1, where a memory cell, e.g., MC2, of the set of NAND strings is connected to a first adjacent word line, e.g., WL2, and another memory cell, e.g., MC0, of the set of NAND strings is connected to a second adjacent word line, e.g., WL0. See FIGS. 3C and 4A.

FIG. 12B depicts a reading process consistent with step 1103 of FIG. 11. The process of FIG. 12B can be performed before or after the process of FIG. 12A. Step 1210 includes selecting a control gate voltage for distinguishing between two adjacent target data states. For example, the two lowest target data states, e.g., the A and B states, may be selected first, in which case the control gate voltage is VrB (see FIGS. 8A and 8G). Step 1211 includes applying the control gate voltage to WLn while sensing whether the selected memory cells are in a conductive state. A decision step 1212 determines whether there is a next control gate voltage to apply. If decision step 1212 is true, step 1210 is performed again by selecting a control gate voltage for distinguishing between the next two adjacent target data states. For example, the next two target data states may be the B and C states, in which case the control gate voltage is VrC. If decision step 1212 is false, the process ends at step 1213.

In one approach, each read process of the additional read processes of FIG. 12B is performed by applying a single control gate read voltage (e.g., VrB, VrC, VrD, VrE, VrF, VrG) to the selected word line while sensing whether the memory cells are in a conductive state.

FIG. 13A depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with one Vth range per data state and four total data states, consistent with step 1101 of FIG. 11. The horizontal axis depicts Vth and the vertical axis depicts a number of memory cells on a logarithmic scale. As mentioned, each Vth range can encompass part of a data state, one data state or multiple data states. In this example, each Vth range encompasses (e.g., spans) exactly one data state, e.g., a range of Vth which is associated with one data state. Specifically, Vth ranges R1, R2, R3 and R4 encompass the Er, A, B and C states, respectively. The Vth ranges may be equal in width, in one approach. Read voltages of VrA, VrB and VrC are used to demarcate between the Er and A states, A and B states, and B and C states, respectively.

FIG. 13B depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with two Vth ranges for the highest data state (C), one Vth range for the next data state (B), and one Vth range for the lowest two data states (Er and A), and four total data states, consistent with step 1101 of FIG. 11. In this example, R4 encompasses a higher portion of the C state (such as a top half), R3 encompasses a lower portion of the C state (such as a bottom half), R2 encompasses the B state and R1 encompasses the Er and A states. The number of voltage ranges is the same as in FIG. 13A, but the voltage ranges are of different widths, with a highest voltage range (R4) having the smallest width and the lowest voltage range (R1) having the largest width. Further, the C state distribution is split into two portions: upper and lower, or highest and next highest. A state could be split into two or more portions. This approach allows the compensation for program disturb to be made with a finer granularity since the Vth of the memory cell which caused the disturb is known more accurately. Also, due to the nonlinearity of the program disturb, the amount of program disturb and the corresponding compensation can differ substantially with a small change in the Vth of the adjacent memory cell, even a change within one data state's Vth distribution. That is, a memory cell on WLn−1 in R**4** may cause greater program disturb than a memory cell on WLn−1 in R**3**. A control gate voltage on WLLn can therefore be compensated more strongly in the former case.

In this example, the Vth range (R**4**) of the unselected memory cell encompasses a highest portion **1302** of a highest data state (C) of the plurality of data states, and a next lower threshold voltage range (R**3**) of the plurality of threshold voltage ranges which is adjacent to the threshold voltage range (R**4**) of the unselected memory cell encompasses a next highest portion (**1301**) of the highest data state. That is, the threshold voltage range of the unselected memory cell is defined with a granularity of less than one data state, so that the threshold voltage range of the unselected memory cell encompasses less than the Vth range of one data state. The Vth range of one data state can be a predetermined value, in one approach.

FIG. **13**B is also an example of the plurality of threshold voltage ranges comprising a highest threshold voltage range (R**4**) and a lowest threshold voltage range (R**1**), where the highest threshold voltage range is smaller than the lowest threshold voltage range. Also, the highest threshold voltage range (R**4**) encompasses a highest data state (C) of the plurality of data states, and the lowest threshold voltage range (R**1**) encompasses a lowest data state (A) of the plurality of data state.

Since the number of voltage ranges is the same as in FIG. **13**A, the number of read voltages and the read time is the same. Read voltages of Vr**1**, Vr**2** and Vr**3** are used to demarcate between R**1** and R**2**, R**2** and R**3**, and R**3** and R**4**, respectively.

FIG. **14**A depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with one Vth range per data state and eight total data states, consistent with step **1101** of FIG. **11**. The horizontal axis depicts Vth and the vertical axis depicts a number of memory cells on a logarithmic scale. In this example, each Vth range encompasses exactly one data state. Specifically, Vth ranges R**1**, R**2**, R**3**, R**4**, R**5**, R**6**, R**7** and R**8** encompass the Er, A, B, C, D, E, F and G states, respectively. The Vth ranges may be equal in width, in one approach. Read voltages of VrA, VrB, VrC, VrD, VrE, VrF and VrG are used to demarcate between the Er and A states, A and B states, B and C states, C and D states, D and E states, E and F states, and F and G states, respectively. The verify levels during programming are VvA, VvB, VvC, VvD, VvE, VvF and VvG.

FIG. **14**B depicts Vth ranges associated with the reading of unselected memory cells on WLn−1, with one Vth range for each of the three highest data states (E, F and G), one Vth range for the next two highest data states (C and D), and one Vth range for the lowest three data states (Er, A and B), and eight total data states, consistent with step **1101** of FIG. **11**. In this example, R**3**, R**4** and R**5** encompass the E, F and G states, respectively. R**2** encompasses the C and D states, e.g., multiple data states of the plurality of target data states.

R**1** encompasses the Er, A and B states. The number of voltage ranges is less than in FIG. **14**A, and the voltage ranges are of different widths, with the highest voltage range (R**5**) having the smallest width and the lowest voltage range (R**1**) having the largest width. This approach allows the compensation for program disturb to be made with a finer granularity for the higher data states which cause the most program disturb, while reducing the total number of Vth ranges. As a result, the number of read processes and the read time is less than in FIG. **14**A. Read voltages of Vr**1**, Vr**2**, Vr**3** and Vr**4** are used to demarcate between R**1** and R**2**, R**2** and R**3**, R**3** and R**4**, and R**4** and R**5** respectively.

This is an example of a plurality of threshold voltage ranges comprising a highest threshold voltage range (R**5**) and a lowest threshold voltage range (R**1**), where the highest threshold voltage range encompasses one number of data states (e.g., one state—the G state) of the plurality of data states, and the lowest threshold voltage range encompasses another number of data states (e.g., three states—the A, B and C states) of the plurality of data states. The one number (1) is smaller than the another number (3).

Various other configurations of the Vth ranges are possible. The Vth levels of the memory cells on WLn−1 are relevant indicators of the amount of program disturb and, in some approaches, the data states of the memory cells on WLn−1 need not be determined.

FIG. **15**A depicts control gate voltages applied to an unselected word line, consistent with step **1101** of FIG. **11**. The horizontal axis depicts time and the vertical axis depicts the control gate voltage on WLn, consistent with FIG. **8**A, for example. Note that it is acceptable for the read of WLn−1 to have a relatively low accuracy since it is not for the purpose of recovering user data. A single voltage can be used to distinguish between each pair of adjacent data states. For example, control gate voltages **1500**, **1501** and **1502** at levels of VrA, VrB and VrC, respectively can be used. A sensing result SR**1**, SR**2**, SR**3** is obtained during each control gate voltage. A sensing result for a memory cell indicates whether the memory cell is in a conductive or non-conductive state, e.g., by a bit value of 0 or 1.

FIG. **15**B depicts control gate voltages applied to a selected word line, consistent with steps **1102** and **1103** of FIG. **11**. The horizontal axis depicts time and the vertical axis depicts the control gate voltage on WLn, consistent with FIG. **8**E, for example. As mentioned in connection with FIG. **12**A, the read operation for distinguishing between the Er and A states for memory cells on WLn involves multiple read processes, one for each data state or Vth range into which the memory cells on WLn−1 are classified. In particular, sensing results SR**1***a*, SR**1***b*, SR**1***c* and SR**1***d* are obtained while control gate voltages **1510**, **1511**, **1512** and **1513** of magnitude VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3** and VrA_C/R**4**, respectively, are applied to WLn. For each memory cell on WLn, one of these sensing results is used while the others are discarded according to the data state or Vth range of the adjacent memory cell on WLn−1.

Subsequently, a sensing result SR**2** is obtained while a control gate voltages **1510** of magnitude VrB is applied to WLn, and a sensing result SR**3** is obtained while a control gate voltages **1520** of magnitude VrC is applied to WLn. Based on three sensing results (SR**2**, SR**3** and one of SR**1***a*, SR**1***b*, SR**1***c* or SR**1***d*), the data state of a memory cell on WLn can be determined.

The read operation thus performs a set of read processes for one memory cell which distinguishes between an erased state and a lowest target data state of a plurality of target data states by applying a plurality of control gate read voltages (**1510**-**1513**) to the one word line while sensing whether the one memory cell is in a conductive state, and selecting a result (SR**1***a*-SR**1***d*) from the sensing based on the Vth range of another memory cell on WLn−1. The read operation further includes additional read processes for the one memory cell for distinguishing between higher target data states (higher than the lowest target data state) of the plurality of target data states by applying additional control gate read voltages (**15010**, **1520**) to the one word line while sensing whether the one memory cell is in a conductive state, wherein each read process of the additional read processes is independent of the Vth range or data state of the another memory cell.

FIG. **15**C depicts a table showing hard bits and soft bits consistent with FIG. **15**B. Each row indicates a voltage on WLn and whether the sensing results provides a hard bit (HB) or soft bit (SB) based on the WLn−1 state of Vth range. Each read operation for a memory cell can result in one hard bit and multiple soft bits. For example, with VWLn=VrA_Er/R**1** (See FIG. **8**E), the sensing result is a hard bit if the WLn−1 state/Vth range is Er/R**1** and a soft bit if the WLn−1 state/Vth range is A/R**2**, B/R**3** or C/R**4**. With VWLn=VrA_A/R**2**, the sensing result is a hard bit if the WLn−1 state/Vth range is A/R**2** and a soft bit if the WLn−1 state/Vth range is Er/R**1**, B/R**3** or C/R**4**. With VWLn=VrA_B/R**3**, the sensing result is a hard bit if the WLn−1 state/Vth range is B/R**3** and a soft bit if the WLn−1 state/Vth range is Er/R**1**, A/R**2** or C/R**4**. With VWLn=VrA_C/R**4**, the sensing result is a hard bit if the WLn−1 state/Vth range is C/R**4** and a soft bit if the WLn−1 state/Vth range is Er/R**1**, A/R**2** or B/R**3**.

The multiple read processes for distinguishing between the Er and A states on WLn can be used in an error correction code (ECC) decoding process. ECC decoding can use probabilistic techniques to improve the accuracy with which data is read from the memory cells. Examples of ECC techniques include Low-Density Parity Check (LDPC) codes, Turbo codes and concatenated codes. LDPC codes are highly efficient linear block codes made from many single parity check (SPC) codes. Turbo codes use iterative soft-decoding and combine two or more relatively simple convolutional codes and an interleaver to produce a block code. Concatenated codes combine algebraic block codes and convolutional codes. These ECC techniques detect hard bits and soft bits to get closer to Shannon limits.

Soft bits can be used to confirm the information obtained from a hard bit to make a read decision more accurate. The soft bits can indicate the shape of the upper or lower tail of a Vth distribution. For example, in FIG. **8**E, consider the hard bit read at VrA_Er/R**1**. The soft bits reads at VrA_C/R**4**, VrA_B/R**3** and VrA_A/R**2** identify the lower tail of the A state Vth distribution **810** to help confirm that the read at VrA_Er/R**1** is between the E state Vth distribution **800** and the A state Vth distribution **810**.

Regarding the hard bit read at VrA_A/R**2**, the soft bit read at VrA_Er/R**1** identifies the upper tail of the Er state Vth distribution **850** and the soft bits reads at VrA_B/R**3** and VrA_C/R**4** identify the lower tail of the A state Vth distribution **810** to help confirm that the read at VrA_A/R**2** is in a valley between the E state Vth distribution **850** and the A state Vth distribution **810**.

Regarding the hard bit read at VrA_B/R**3**, the soft bit reads at VrA_Er/R**1** and VrA_A/R**2** identify the upper tail of the Er state Vth distribution **851** and the soft bits read at VrA_C/R**4** identifies the lower tail of the A state Vth distribution **810** to help confirm that the read at VrA_B/R**3** is in a valley between the E state Vth distribution **851** and the A state Vth distribution **810**.

Regarding the hard bit read at VrA_C/R**4**, the soft bit reads at VrA_Er/R**1**, VrA_A/R**2** and VrA_B/R**3** identify the upper tail of the Er state Vth distribution **852** to help confirm that the read at VrA_C/R**4** is in a valley between the E state Vth distribution **852** and the A state Vth distribution **810**.

Thus, we can have reads which serve the double purpose of providing soft bits which assist in decoding data as well as compensate for disturb to the Er state memory cells. In this case, no additional read time is consumed compared to the case where the soft bits were obtained to assist in decoding data but not to compensate for disturb to the Er state memory cells. Additional reads can be performed to obtain additional soft bits. Moreover, reads which provide soft bits can also be used for distinguishing between the target data states.

Accordingly, it can be seen that, in one embodiment, a method for operating a memory device comprises: (1) programming memory cells in a set of word lines in a word line programming order, the set of word lines comprise one word line (WLn), a first adjacent word line (WLn+1) which is adjacent to the one word line and after the one word line in the word line programming order and a second adjacent word line (WLn−1) which is adjacent to the one word line and before the one word line in the word line programming order, wherein one memory cell (MC**1**) is connected to the one word line, another memory cell (MC**0**) is connected to the second adjacent word line, the one memory cell and the another memory cell are arranged along a continuous charge-trapping film, the programming of the memory cells comprises applying a plurality of program pulses to the one word line while a first pass voltage is applied to the first adjacent word line and a second pass voltage is applied to the second adjacent word line, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses; and (2) subsequently, in response to a read command involving the one memory cell: (a) reading the another memory cell to determine a threshold voltage range among a plurality of threshold voltage ranges (R**1**-R**8**) of the another memory cell, and (b) after the reading of the another memory cell: (i) performing a set of read processes for the one memory cell which distinguishes between an erased state and a lowest target data state of a plurality of target data states by applying a plurality of control gate read voltages (VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3**, VrA_C/R**4**; VrA_Er,A,B, VrA_C,D, VrA_E, VrA_F, VrA_G) to the one word line while sensing whether the one memory cell is in a conductive state, and selecting a result (SR**1**_a_-SR**1**_d_) from one read process of the set of read processes based on the threshold voltage range of the another memory cell, and (ii) performing additional read processes for the one memory cell for distinguishing between higher target data states of the plurality of target data states, wherein each read process of the additional read processes is independent of the threshold voltage range of the another memory cell.

In another embodiment, a memory device comprises: (1) a set of NAND strings, each NAND string comprising memory cells and a charge-trapping film which extends continuously in the NAND string; (2) a set of word lines connected to the set of NAND strings; and (3) a control circuit. The control circuit is configured to: (a) program memory cells in the set of word lines in a word line programming order, starting from a source side of the set of NAND strings and proceeding toward a drain side of the set of NAND strings, wherein during programming of memory cells of one word line (WLn) of the set of word lines, a plurality of program pulses are applied to the one word line while a first pass voltage is applied to a drain-side word line (WLn+1) in the set of word lines which is adjacent to and on a drain side of the one word line, and a second pass voltage is applied to a source-side word line (WLn−1) in the set of word lines which is adjacent to and on a source side of the one word line, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses, (b) read a memory cell (MC**0**) which is connected to the source side word line (WLn−1) to determine a target data state among a plurality of target data states of the memory cell which is connected to the source side word line, (c) perform a set of read processes for one memory cell (MC**1**) which is connected to the one word line, the set of read processes for the one memory cell distinguishes between an erased state and a lowest target data state of a plurality of target data states by applying a plurality of control gate read voltages (VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3**, VrA_C/R**4**; VrA_Er,A,B, VrA_C,D, VrA_E, VrA_F, VrA_G) to the one word line while sensing whether the one memory cell is in a conductive state, and selecting a result (SR**1***a*-SR**1***d*) from one read process of the set of read processes based on the target data state of the memory cell which is connected to the source side word line, and (d) perform additional read processes for the one memory cell for distinguishing between higher target data states of the plurality of target data states, wherein each read process of the additional read processes is independent of the target data state of the memory cell which is connected to the source side word line.

In another embodiment, a memory device comprises: (1) a string of memory cells connected in series and extending along a continuous charge-trapping film; and (2) a control circuit. The control circuit: (a) programs the memory cells in a programming order, starting from a source side of the string and proceeding toward a drain side of the string, wherein during programming of one memory cell (MC**1**) of the string, a plurality of program pulses are applied to the one memory cell while a first pass voltage is applied to a drain-side memory cell (MC**2**) in the string which is adjacent to and on a drain side of the one memory cell, a second pass voltage is applied to a source-side memory cell (MC**0**) in the string which is adjacent to and on a source side of the one memory cell, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses, (b) reads the source-side memory cell to determine a measure of a threshold voltage of the source-side memory cell, the source-side memory cell is adjacent to the one memory cell and was programmed before the one memory cell according to the programming order, (c) reads the one memory cell to distinguish between an erased state and a lowest target data state of a plurality of target data states using a plurality of control gate read voltages (VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3**, VrA_C/R**4**; VrA_Er,A,B, VrA_C,D, VrA_E, VrA_F, VrA_G), (d) selects a result (SR**1***a*-SR**1***d*) from the read of the one memory cell to distinguish between the erased state and the lowest target data state, based on the measure of the threshold voltage of the source-side memory cell, and (e) reads the one memory cell to distinguish between higher target data states of the plurality of target data states independently of the measure of the threshold voltage of the source-side memory cell.

In another embodiment, a memory device comprises: a set of NAND strings, each NAND string comprising memory cells and a charge-trapping film which extends continuously in the NAND string; a set of word lines connected to the set of NAND strings, the set of word lines comprise one word line (WLn), a first adjacent word line (WLn+1) which is adjacent to the one word line and after the one word line in a word line programming order and a second adjacent word line (WLn−1) which is adjacent to the one word line and before the one word line in the word line programming order, wherein one memory cell (MC**1**) of the set of NAND strings is connected to the one word line (WL**1**) and another memory cell (MC**0**) of the set of NAND strings is connected to the second adjacent word line; and a control circuit. The control circuit is configured to: program memory cells of the set of NAND strings in the word line programming order, wherein during programming of memory cells of the one word line (WLn), a plurality of program pulses are applied to the one word line while a first pass voltage is applied to the first adjacent word line and a second pass voltage is applied to the second adjacent word line, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses.

Subsequently, in response to a read command involving the one memory cell (MC**1**), the control circuit is configured to: read the another memory cell (MC**0**) to determine a threshold voltage range among a plurality of threshold voltage ranges (R**1**-R**8**) of the another memory cell, and after the read of the another memory cell: perform a set of read processes for the one memory cell which distinguishes between an erased state and a lowest target data state of a plurality of target data states by applying a plurality of control gate read voltages (VrA_Er/R**1**, VrA_A/R**2**, VrA_B/R**3**, VrA_C/R**4**) to the one word line while sensing whether the one memory cell is in a conductive state, and selecting a result (SR**1***a*-SR**1***d*) from one read process of the set of read processes based on the threshold voltage range of the another memory cell, and perform additional read processes for the one memory cell for distinguishing between higher target data states of the plurality of target data states, wherein each read process of the additional read processes is independent of the threshold voltage range of the another memory cell.

The foregoing detailed description of the invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. The described embodiments were chosen in order to best explain the principles of the invention and its practical application, to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the claims appended hereto.

What is claimed is:

1. A method for operating a memory device, comprising:

programming memory cells in a set of word lines in a word line programming order, the set of word lines comprise one word line, a first adjacent word line which is adjacent to the one word line and after the one word line in the word line programming order and a second adjacent word line which is adjacent to the one word line and before the one word line in the word line programming order, wherein one memory cell is connected to the one word line, another memory cell is connected to the second adjacent word line, the one memory cell and the another memory cell are arranged along a continuous charge-trapping film, the programming of the memory cells comprises applying a plurality of program pulses to the one word line while a first pass voltage is applied to the first adjacent word line and a second pass voltage is applied to the second adjacent word line, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses; and

subsequently, in response to a read command involving the one memory cell:

reading the another memory cell to determine a threshold voltage range among a plurality of threshold voltage ranges of the another memory cell, and

after the reading of the another memory cell:

performing a set of read processes for the one memory cell which distinguishes between an erased state and a lowest target data state of a plurality of target data states by applying a plurality of control gate read voltages to the one word line while sensing whether the one memory cell is in a conductive state, and selecting a result from one read process of the set of read processes based on the threshold voltage range of the another memory cell, and

performing additional read processes for the one memory cell for distinguishing between higher target data states of the plurality of target data states,

wherein each read process of the additional read processes is independent of the threshold voltage range of the another memory cell.

2. The method of claim 1, wherein:

each read process of the additional read processes is performed by applying a single control gate read voltage to the one word line while sensing whether the one memory cell is in a conductive state.

3. The method of claim 1, wherein:

the set of read processes for the one memory cell use an error correction code decoding process based on each read process of the set of read processes.

4. The method of claim 1, wherein:

the result from the one read process is used to provide a hard bit in an error correction code decoding process; and

a result from another read process of the set of read processes is used to provide a soft bit in the error correction code decoding process.

5. The method of claim 1, wherein:

the threshold voltage range of the another memory cell encompasses one or more target data states of the plurality of target data states.

6. The method of claim 1, wherein:

the threshold voltage range of the another memory cell encompasses multiple data states of the plurality of target data states.

7. The method of claim 1, wherein:

the threshold voltage range of the another memory cell encompasses a highest portion of a highest target data state of the plurality of target data states; and

a next lower threshold voltage range of the plurality of threshold voltage ranges which is adjacent to the threshold voltage range of the another memory cell encompasses a next highest portion of the highest target data state.

8. The method of claim 1, wherein:

the one memory cell and the another memory cell are arranged along in a three-dimensional stacked memory structure comprising alternating conductive layers and dielectric layers;

the first adjacent word line is provided by one of the conductive layers; and

the second adjacent word line is provided by another of the conductive layers.

9. The method of claim 1, wherein:

a difference by which the first pass voltage exceeds the second pass voltage during the one or more initial program pulses is at least as high as a difference between a verify voltage of a highest target data state of the plurality of target data states and a verify voltage of the lowest target data state of the plurality of target data states during the programming of the memory cells.

10. The method of claim 9, wherein:

the first pass voltage and a difference by which the first pass voltage exceeds the second pass voltage start to become progressively smaller during the programming of the memory cells when a determination is made that a specified portion of memory cells which are to be programmed to the lowest target data state of the plurality of target data states have passed a verify test.

11. A memory device, comprising:

a set of NAND strings, each NAND string comprising memory cells and a charge-trapping film which extends continuously in the NAND string;

a set of word lines connected to the set of NAND strings; and

a control circuit, the control circuit is configured to:

program memory cells in the set of word lines in a word line programming order, starting from a source side of the set of NAND strings and proceeding toward a drain side of the set of NAND strings, wherein during programming of memory cells of one word line of the set of word lines, a plurality of program pulses are applied to the one word line while a first pass voltage is applied to a drain-side word line in the set of word lines which is adjacent to and on a drain side of the one word line, and a second pass voltage is applied to a source-side word line in the set of word lines which is adjacent to and on a source side of the one word line, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses,

read a memory cell which is connected to the source side word line to determine a target data state among a plurality of target data states of the memory cell which is connected to the source side word line,

perform a set of read processes for one memory cell which is connected to the one word line, the set of read processes for the one memory cell distinguishes between an erased state and a lowest target data state of a plurality of target data states by applying a plurality of control gate read voltages to the one word line while sensing whether the one memory cell is in a conductive state, and selecting a result from one read process of the set of read processes based on the target data state of the memory cell which is connected to the source side word line, and

perform additional read processes for the one memory cell for distinguishing between higher target data states of the plurality of target data states, wherein each read process of the additional read processes is independent of the target data state of the memory cell which is connected to the source side word line.

12. The memory device of claim 11, wherein:

the read of the memory cell which is connected to the source side word line is performed in response to a read command involving the one memory cell.

13. The memory device of claim 11, wherein:

sensing circuitry which performs the set of read processes and the additional read process is connected to a drain side of the one memory cell.

14. The memory device of claim 11, wherein:

the set of NAND strings is in a three-dimensional stacked memory structure comprising alternating conductive layers and dielectric layers;

the drain-side word line is provided by one of the conductive layers; and

the source-side word line is provided by another of the conductive layers.

15. The memory device of claim 11, wherein:

the result from the one read process is used to provide a hard bit in an error correction code decoding process; and a result from another read process of the set of read processes is used to provide a soft bit in the error correction code decoding process.

16. A memory device, comprising:

a string of memory cells connected in series and extending along a continuous charge-trapping film; and

a control circuit, the control circuit:

programs the memory cells in a programming order, starting from a source side of the string and proceeding toward a drain side of the string, wherein during programming of one memory cell of the string, a plurality of program pulses are applied to the one memory cell while a first pass voltage is applied to a drain-side memory cell in the string which is adjacent to and on a drain side of the one memory cell, a second pass voltage is applied to a source-side memory cell in the string which is adjacent to and on a source side of the one memory cell, and the first pass voltage is higher than the second pass voltage during one or more initial program pulses of the plurality of program pulses,

reads the source-side memory cell to determine a measure of a threshold voltage of the source-side memory cell, the source-side memory cell is adjacent to the one memory cell and was programmed before the one memory cell according to the programming order,

reads the one memory cell to distinguish between an erased state and a lowest target data state of a plurality of target data states using a plurality of control gate read voltages,

selects a result from the read of the one memory cell to distinguish between the erased state and the lowest target data state, based on the measure of the threshold voltage of the source-side memory cell, and

reads the one memory cell to distinguish between higher target data states of the plurality of target data states independently of the measure of the threshold voltage of the source-side memory cell.

17. The memory device of claim 16, wherein:

the one memory cell and the source-side memory cell are in a common NAND string in a three-dimensional structure; and

the three-dimensional structure comprises alternating conductive layers and dielectric layers.

18. The memory device of claim 16, wherein:

the control circuit reads the one memory cell to distinguish between the higher target data states of the plurality of target data states is by performing multiple reads of the one memory cell, each read of the multiple reads involves applying a single control gate read voltage to the one memory cell while sensing whether the one memory cell is in a conductive state.

19. The memory device of claim 16, wherein:

the result from the read of the one memory cell to distinguish between the erased state and the lowest target data state based on the measure of the threshold voltage of the source-side memory cell is used to provide a hard bit in an error correction code decoding process; and

a result from the read of the one memory cell to distinguish between the higher target data states is used to provide a soft bit in the error correction code decoding process.

20. The memory device of claim 16, wherein:

the measure of the threshold voltage of the source-side memory cell encompasses one or more target data states of the plurality of target data states.

21. The memory device of claim 16, wherein:

the first pass voltage and a difference by which the first pass voltage exceeds the second pass voltage start to become progressively smaller during the programming of the memory cells when a determination is made that a specified portion of memory cells which are to be programmed to the lowest target data state of the plurality of target data states have passed a verify test.

* * * * *